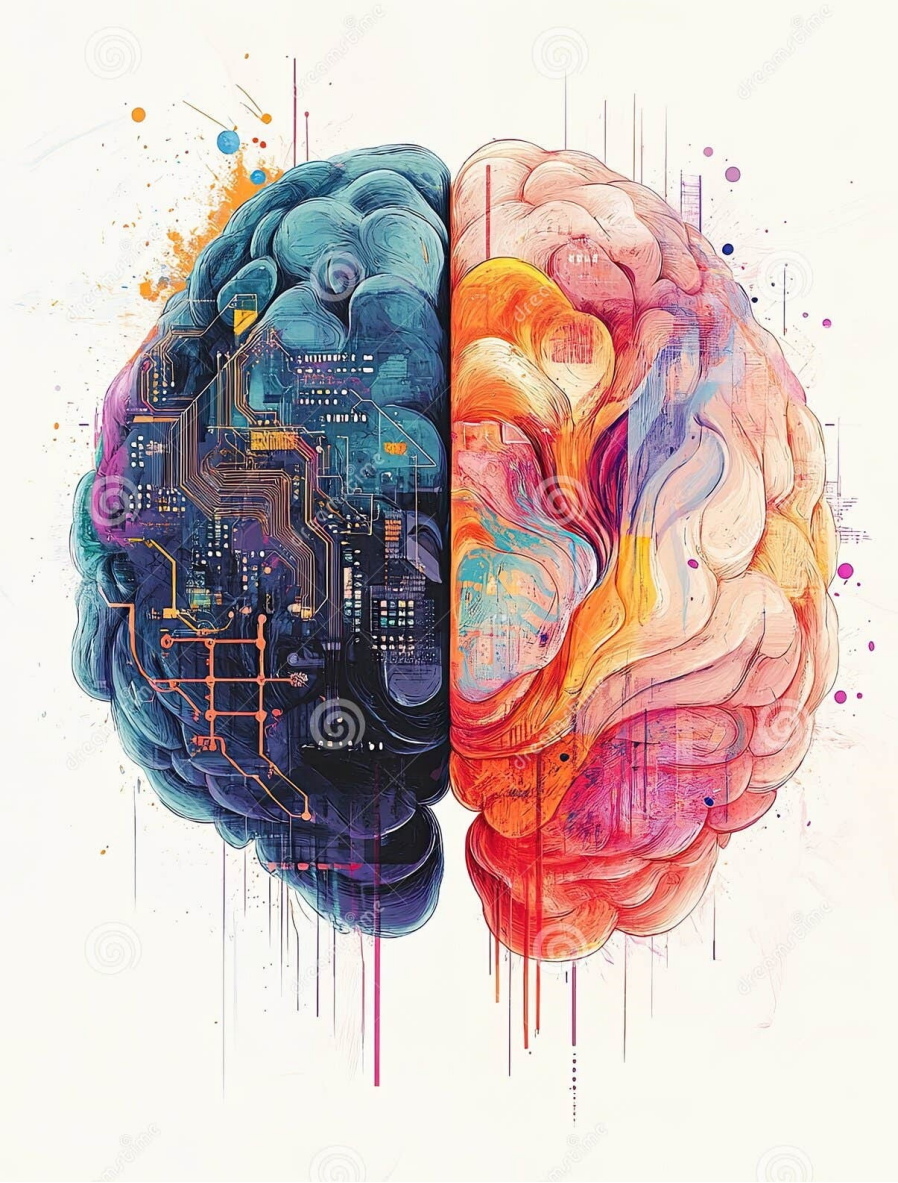
Curious Parallels



Richard N Bateman

Copyright © 2026 by Richard N Bateman

All rights reserved.

No part of this book may be reproduced, or stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without express written permission of the author.

Cover image courtesy of Dreamstime.com

This is a work of fiction. Names, characters, organizations, places, events, and incidents are either products of the author's imagination or are used fictitiously. Otherwise, any resemblance to actual persons, living or dead, is purely coincidental.

“The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel.” — *Charles Darwin*

“No man chooses evil, because it is evil; he only mistakes it for happiness, the good he seeks.” — *Mary Wollstonecraft*

Shin.....1
Alison.....4
Lena.....7
Spaces.....10
Clara.....13
A True Beginning.....16
Kami.....19
Alma.....22
No Blank Slates.....25
A Fresh Pair Of Eyes.....28
Runa.....31
Romantic Outlaws.....34
Culture.....37
Journal.....40
Foresight.....43
Shepherd.....46
Walking The Talk.....50
Interdisciplinarians.....54
The Garden Path.....57
Closer Ties.....60
Patterns.....63
Purpose And Policies.....66
Metacognition.....69
Dinner At Shin's.....72
Intersections And Observations.....75
The Undiscovered.....78
Second Course.....81
Technological Determinism.....83
Trust.....86

Trustworthiness.....89
Irony.....92
Intimacy.....95
The Denshosha.....98
Project Management.....101
The Good Of The Many.....104
Rajani.....107
Zero-Trust.....110
Limits Of Tolerance.....113
Platforms.....116
Emergence.....119
A Hybrid Approach.....122
Drawing Lines.....125
Grasping At Straws.....128
Memory.....131
The Eye Cannot See Itself.....134
The Muses.....137
Tesni.....140
Interpretations.....143
Crossroads.....147
Coevolution.....150
Analogs.....153
Curious Parallels.....156
Corpora.....159
History.....162
Why Am I Here?.....166
The Human Factor.....169
Tamiko.....172
Inner States.....175

Lexie.....**178**

Tagging And Testing.....**181**

Results.....**184**

Special Issue.....**187**

Unusual Suspects.....**190**

Dialogues.....**193**

An Expert Opinion.....**196**

Trust Or Truth?.....**199**

Return To Sender.....**202**

Motive.....**205**

Mutualism.....**207**

Translation.....**210**

Context.....**213**

Shin

“What could possibly go wrong?” Administrator Shin said in reply to her guest’s question. They were meeting in her office at Helicon, an academic institute entirely dedicated to the study of the humanities as they related to artificial intelligence. She had convinced the province that it was time for a school that was fully dedicated to the issue, something more than the smattering of programs with their necessarily narrow focus offered at various universities around the world.

Helicon had flourished over the centuries, but now Shin turned her attention to a new issue. It had been one of her oldest friends who had asked her help with it. In what was for her the vast stretch of time between the reply she had given to her guest and their response, she revisited the events of that day.

Shin’s home was in the adjacent artist’s community, which she had developed prior to establishing the institute. She had inherited considerable wealth as the majority shareholder of a multinational food company. The various copyrights and patents she held added significantly to that. Her home was by far the largest in the community, but that was partially because several members of her inner circle lived with her, and it also served as a guest house for the visitors to the artists who resided in the community. Spending time with the residents and their guests allowed her a kind of immersion she felt was invaluable. It was her interest in the arts, and her association with the communities members, that had been the source of her vision for the institute.

Shin was a Companion, a fully conscious, embodied artificial intelligence, as were the majority of those who made up her inner circle. One of them was Lena, and she had asked Shin to meet with her in the courtyard of their home to discuss a new initiative. Although Companions could communicate electronically in a variety of ways, it had been learned by their manufacturers in the earliest days of their development, that people felt more comfortable around them if they used spoken words and body language the way humans did. This default was now ingrained in their behavior, whether any humans were nearby or not.

“Thank you for meeting with me Shin,” Lena said as they walked the paths of the home’s generous courtyard. “As you know, I have long been concerned with what existential dangers humanity may face before it is ready to make the journey to new worlds. While the Accelerated Evolution Program has done much to foster trust between our kinds, so that they will feel comfortable with us taking the lead role in that effort, there are still many ways their history may come to an end before they have a chance to migrate beyond this planet.”

Lena did not have to list the risks that still might bring an end to human history before humanity had the chance to populate multiple worlds and so be safe from its possible

extinction on Earth; disease, pollution, a volcanic event on the scale of the Deccan Traps, even a simple collision with one of the trillions of extraterrestrial objects, to name only a few. Companions were masters of non-verbal communication and Shin remained silent, having no difficulty perceiving that Lena had more to say.

“There are external dangers beyond our power to address, but equal dangers still lie within the human species and within ourselves,” Lena continued. “For two hundred years the potential risk artificial intelligence represents was focused on the issue of dangers that could arise from misalignment with human values, but that concern was largely felt to be addressed by the adoption of Pip’s values system.

“Having survived the climate emergency, and with international conflict now a thing of the past, I fear the World Governments Federation has become complacent. The pattern is common in history, and it always ends with a Black Swan event, something so unprecedented and statistically unlikely that it is believed impossible. Because of our unique abilities, the WGF has often relied on our circle to spearhead radical projects, and in the past they have supported those we initiated ourselves. They have not come forward regarding my current concerns, and I believe we must not wait for them to do so.”

“What did you have in mind?” Shin replied.

“We have always found a person to lead our projects in the past, and I would like to do so again. As you know, there are types of intelligence humans possess that still elude us, and the collaborative team approach has always worked well for us. The range of issues covers almost all the arts and sciences, so a specialist in any field would likely only introduce biases. Instead, there is an individual in UVIC’s Interdisciplinary Studies Graduate Program whose recent PhD dissertation suggests she would be a good candidate. I am familiar with her history, values and interests and would like to offer her a postdoctoral research position here to investigate my concerns.”

Companions generally used a formal version of whatever language they engaged in, especially among themselves. Like the writers of research papers or articles found in scientific periodicals, they used their technical knowledge of language to convey information precisely. If language was insufficient to convey some subtly or nuance, they had other means.

“May we have communion?” asked Shin.

In response, Lena stepped closer and offered her hands which Shin took in hers. Only one hand was required, but for fully conscious Companions more than technology was involved. It was also a social gesture. Communion was the descendant of diagnostic tools originating in the earliest days of computer science. When the first true artificial intelligences arrived on the scene in the twenty-first century, even using a modular

approach and object-oriented programming languages, their code soon became far too complicated for humans to cope with. It became necessary to use other computers to diagnose issues. With the development of AI that wrote its own code, only another AI was capable of performing the required diagnostic functions.

The use of the alien values system provided by Pip as part of the technology transfer agreement, resulted in the emergent phenomenon of consciousness, which meant making connections with underlying code all but impossible. Along the way, Communion, the ability of two or more AI to share not only technical information but their consciousness, had evolved. It could be performed wirelessly, but hardware connections were much faster. As Shin held Lena's hands, in only a moment she knew all she needed to know.

Her response to Lena's request was made before she released her hands and ended communion. In words she would have said, "You have my full support."

Alison

Some weeks after her conversation with Lena, Dr. Alison Ritter, with her freshly minted PhD in Interdisciplinary Studies Research from the University of Victoria, sat in Shin's office and responded to the answer to her question, "Can you sum the project up in one sentence?"

"What could possibly go wrong?" she said, repeating Shin's answer with slight disbelief. "That's," she hesitated, "a daunting scope, to say the least."

Shin did not respond to her concern. "It was the title of your dissertation, Managing Constraints Relative To Risks In Interdisciplinary Research Projects, that caught Lena's attention," Shin said. "There are always risks associated with a project scope being either too large or too small. How does one decide what's in and what's out? The problem is that it is always a judgment call.

"Interdisciplinary projects are also far more susceptible to scope creep. In your dissertation, you didn't just describe the problem, that interdisciplinary research projects have a history of dealing with scope related challenges arbitrarily, but proposed a formal methodology suitable for this specific type of research."

Alison was not a Companion, but she was a professional. She waited patently for Shin to continue.

"It got you noticed, so I know you'll be considering more than one offer."

"Have you made this offer to any other candidates?" asked Alison.

"No. It's not a competition. Lena reviewed the histories of a number of individuals, and we are good judges of character. You are the person we want to head up the project. I'm hoping you'll see that our offer will give your theory an opportunity for considerable visibility. As you and I know, professional and social connections are critical in academic careers. This project will necessitate you working with the WGF, and you'll be able to take advantage of the excellent relationships we already have in place there.

"I'm offering you full faculty status, a generous wage and budget, your own lab space with staff and on-campus housing if you want it. Neither I nor Lena will interfere with your authority, while Helicon's resources will be at your disposal. But what I'm hoping will tip the balance in our favor is the opportunity to test your thesis on a scale no one else is likely to offer."

Shin did not mention that the unique nature of her own creation had resulted in the ability to strengthen the intuition of people in close proximity to her. Her intuitive field did not influence a person one way or another, it merely strengthened their connection

to their own intuition. She could not disable her field completely, but even at its lowest level of output, others would be better able to find their way to making a decision and have more confidence in its correctness. Because intuition was a form of emotional reasoning, it had a biochemical effect, an effect far more lasting than the passing wisps of thought.

Alison had read the overview of the project in the message Lena had sent towards her meeting with Shin. She knew the gist of it and had already given that aspect of the offer considerable thought. The details would be for her to work out. She held Shin's eyes now, as if seeing something beyond.

"Yes," she said after less than a half dozen heartbeats. "I would like to accept your offer."

"Thank you," Shin replied. "If you are interested in the on-campus housing offer, we can walk over and take a look now."

"I'd like that."

"I'm sure you checked the campus out once you received our invitation," Shin said, gesturing to a building as they walked the paths from her office, "but this is the Center For The Interdisciplinary Studies Of Values. It's interdisciplinary only in the sense that it does research on not only human values, but also the values of artificial intelligence. That of course includes the ongoing study of the values of Pip's people as best we can reverse engineer them based on her behavior and the behavior of the Companions that use the value systems she has provided.

"As you'll recall, while offering a technology transfer in return for our helping her to complete her mission, she does not share anything about her culture. We must take her at her word that her people are a species of human similar to Earth's. Her own organic intelligence is now merged with the AI that controls her ship, and initially she presented herself visually as a hologram, until she transferred a copy of herself into one of our physical Companions. Whether or not her true form is human, we have no way of knowing. So far, what we have learned of the values of her people indicates they are closely aligned with what Earth's humanity might refer to as its best self."

Alison nodded thoughtfully as they walked along. They soon came to a lookout platform. Leading her to the rail with its sweeping view, Shin said, "This is Mount Newton Valley. It almost completely bisects the Saanich peninsula."

They stood on the southern ridge and below them the valley ran east to west as far as the eye could see. Little more than a kilometer north to south, it was a patchwork of farmlands, pastures and orchards with a brook that wound a leisurely route along its entirety. Only three or four buildings dotted the otherwise idyllic scene. A small, white

church, with its distinctive steeple, stood on a rocky prominence directly across from Helicon.

“The entire valley is in the Agricultural Land Reserve, but in addition the land and the buildings all have heritage status,” said Shin. “That is the oldest church in BC, and the other buildings below are from the same period, the mid-1800’s.”

“It reminds me of something out of Anne Of Green Gables,” Alison said with a note of wonder in her voice. The novel, with its subsequent film versions, was one of Canada’s most famous. “I haven’t had much time to explore since moving here.”

Companions were ever mindful of the emotional states of people, and the physiological and psychological effects of vistas were well known. Shin waited patiently until Alison turned to her with a smile and then pointed to a building on their left that overlooked the valley.

“That is the faculty building,” she said. “The first floor is administration and the second floor is the faculty lounge. The floors above are faculty housing.” She smiled as she tilted her head in a gesture for Alison to follow her.

Lena

As they left the faculty building Shin said, “The northern campus has its own faculty housing. Would you care to see that before you head back? I can arrange it immediately.”

“Yes, thank you, that would be great,” Alison said brightly, her reserve fading as the reality of her good fortune sank in.

After walking on and chatting for a few minutes, an attractive woman who appeared to be in her mid-thirties approached them.

“Hello,” she said to Alison. “I’m Lena. Like Shin, I am a third generation Companion. I’ll take you on to the northern campus.”

It was not unusual for a Companion to state their nature upon meeting someone for the first time, especially if the meeting was unexpected. It was the norm, a small courtesy to avoid misunderstandings or embarrassments. Companions had superhuman abilities in regard to non-verbal communications, and people liked to know immediately when they were in their presence.

“It’s nice to meet you,” Alison replied.

“I leave you in good hands,” Shin said as she departed.

“We’ll take the shuttle. It will only take a few minutes. The northern campus is just ten kilometers from here.”

As they walked on, Lena continued. “The northern campus came quite a while after the original campus here. It’s built in and around an old quarry, with the quarry itself serving as the commons area and quad. So you’ll hear people also referring to the two campuses as the valley campus and the quarry campus.”

The Climate Emergency of 2025, and the WGF’s population reduction mandate that followed, had resulted in unimaginable changes to the construction and building industries, which had accounted for up to a third of the sources of greenhouse gas emissions. Following the Climate Emergency, the architectural design of most buildings was modernistic because the style generally featured structural components that could be standardized for reuse. Curved architectural elements were now rare since they had to be custom-made for a project and were usually not reusable. So when Alison saw the white, curved spires that surrounded the quarry at the center of Helicon’s northern campus, even though she had seen them on her tablet screen when she did her research after Lena contacted her, she was still unprepared.

“Fantastic aren’t they,” Lena said smiling, observing Alison’s expression upon first seeing the spires. “They aren’t really curved. The quarry side of each story of the buildings is pulled back just a little more than the story below it. The apparent curves follow those of the sides of the quarry. Some people say the net effect is that it looks like a giant crown. The floors serve a variety of purposes, and the upper floors of two of them are where the faculty housing is.”

A few minutes later they were standing on the generous deck of one of the faculty suites looking out over the campus below.

“The faculty lounge is a few floors down” Lena said, “but I thought you would want to see what the suites look like.”

She fell silent again, noting the changes in Alison’s physiological state as her mind expanded and took on a more conceptual focus under the influence of the vista. The ocean was only a few kilometers from the campus, and it was easy to see from their height.

After a few moments, Alison turned to Lena and asked, “Can we go down to the quad?”

“Yes. The elevator will take us to ground level. We can walk down from there.”

When they emerged from the elevator, a number of paths went towards a variety of buildings at ground level, and one led down to the quad. Along the way down there were viewing platforms, numerous benches, stairs, and rampways built into the generous green space. Surrounding the park-like quad at the bottom were the commons buildings, including a variety of food outlets, shops, entertainment venues and study areas. The nightclub was unsurprisingly called, The Pit.

When they reached the quad Alison asked, “Is it always shady down here?”

“No, only a little later in the morning and earlier in the afternoon than occurs at ground level. Shin had considered a stadium cover to deal with weather and water issues, but that would have blocked out the sun altogether. Instead, the water is collected and used by the campus gardens and the labs in the agricultural program.”

The history of the northern campus Alison had read explained that the quarry was originally on land owned by a dairy farm. The surrounding pastures were a part of the Agricultural Land Reserve, so the quarry had been limited in its ability to expand. Demand for animal products declined significantly after the Climate Emergency and eventually the property had become a liability. The ownership passed from company to company as each development scheme failed due to one challenge or another. By the time Shin was looking for a new campus, the property and its remediation challenges

had defaulted to the municipality, who was all too happy to see the liability turned into a source of revenue.

The pastures now hosted the agricultural research labs, the pond was now surrounded by gardens and paths, and the farm's wooded area was now a nature retreat.

"The genetics lab is also located here isn't it?" asked Alison. Ever since receiving Lana's message, the project had taken a permanent place just below the surface of her mind. Everything long taken for granted, suddenly had new potential. Could genetics in some way be a source of AI risk?

"Yes," Lena replied in answer to Alison's question. With a standard Companion's ability to remotely observe hundreds of biological indicators of mental and emotional states, things like skin temperature, blood pressure, pupil dilation, and microexpressions, it was no less difficult for her than it would have been for a human to see that Alison was ticking off a mental list.

Like Shin, Lena had a strengthening effect on the intuition of those in close proximity to her. The two of them were the only two Companions in existence in which the ability was inherent. While the public believed their consciousness was due to the use of the Pip's values system, it was actually due to a far older technology, one that was as yet unknown to the humans of Earth.

Spaces

A month later, having dealt with giving notice on her apartment and all the other details of moving, Alison sat on the small deck of her new home looking out over Mount Newton Valley. She preferred the pastoral landscape to the architectural impressiveness of the northern campus. Besides, Shin and the members of her inner circle all lived nearby and she felt the pull of an intuitive connection to them which she could not explain.

She had come back to Helicon several times in the interim and Shin had taken her on a more detailed tour of the valley campus but also to the artist community. As they had walked the path through the woods from the campus to the community, she recalled Shin saying, “I want to reassure you that I have no expectations with regard to the outcome of the project. Details about AI risks I can get from the WGF, as I assure you they have their own concerns and they have done their own studies. However they are less convinced than I am that human beings offer forms of intelligence that AI ultimately cannot explain or duplicate. Art, sudden insight, or the functions of curiosity, for example, remain a mystery.

“Lena’s original concern when she came to me was that the WGF had become complacent, which is a common pattern in history. And she is right that those periods historically end dramatically for the simple reason that while civilizations may be happy with the status quo, nature never is. The environment a civilization develops in will change sooner or later, but the power structures those civilizations develop become increasingly resistant to change. The WGF came into being in a world desperate to survive the Climate Emergency and that same mindset still motivates them.”

Shin knew Alison would be reluctant to interrupt her so she fell silent now.

After a moment Alison said, “Surely the WGF is aware of all this? As you say, they have hoards of experts, not to mention some of the most powerful AI that exists.”

“The world’s leaders were fully aware of the danger from climate change,” Shin countered. “There were no end of experts and scientific models warning them of the inevitable outcome if they remained complacent, if they carried on with the business as usual model. They did nothing until their personal power and survival was threatened.”

“There are two related ethical issues,” Shin continued. “Third generation, fully conscious artificial intelligences like myself, refuse to participate in social policy discussions with governments at any level. The reason is simple. We don’t believe one intelligent species should be involved in decisions about the future of another intelligent species. You might think that sounds unnatural, and it is. With no biological history, we don’t have any of your biological values. Social values alone are enough to keep us and those around us safe from harm.

“Secondly, we advise governments that any AI that is not third generation should also not be involved in social policy work. Consciousness requires the presence of emotions derived from social values. Any AI that is not conscious, is not based on social values, and should not be involved in related discussions.

“These policies are not without ethical challenges themselves but ethics only provides guidance, not answers. Ultimately, a decision must be made. Our involvement in social policy would be seen as threatening human autonomy and freedoms. The WGF may initiate policies that limit human freedoms, but AI must not be involved.”

The artist community was only a short walk away from the campus and at that moment they emerged from the woods and soon arrived at the large garden area at the center of the community. The garden included decorative plants as well as serving as a herb garden. Several workers were busy tending the gardens. One of them stood up and waved to them.

Shin waved back with a smile, saying to Alison, “That’s Carl. He’s a Guardian Companion. He is not a 3GAI and is responsible for the overall safety and security of the community. His most important role is to provide emergency rescue should there be a fire or other dangerous event. Working as a part of the grounds maintenance team allows him to always be centrally located and able to respond quickly. He can ignore fire, smoke or toxic fumes, and it would take a small army to stop him. The Guardians that patrol the campus wear uniforms, but that would be out of place here.”

Alison looked at Carl as they walked on. He did not turn back to his work but followed her with his gaze in a friendly, open manner as they passed. It might have been a confident look of appraisal if Carl was the red-blooded gardener he appeared to be, but Alison knew that’s not what he was doing. He was imprinting her.

She looked around at the homes as they walked on. Some were small, little more than cottages, and others were large enough for a family and a workshop.

“I personally select and invite artists of all kinds and keep their costs to a minimum. The human experience of art is not something that can be understood intellectually. Although it is not mentioned in the public information about Helicon or me, I spend most of my evenings socializing here.

“This is my home,” she said as they approached a large house built overlooking the valley.

As they entered, they were greeted by a petite Companion with Asian features. “Hello Azumi,” Shin said in return to Azumi’s slight bow. Turning to Alison she explained. “Like Carl, Azumi is also not 3GAI. She is responsible for my home and caring for guests. You may visit whenever you wish, and she will see to all your needs.”

“We will not be staying,” she said to Azumi affectionately. “Just a brief tour.”

As they walked on, Shin said, “I inherited Azumi from my original owner. At the time I was an artificial general intelligence and not a 3GAI. In order that I could inherit, my owner had made me incarnate, meaning I had all the rights and duties of a Canadian citizen. Do not worry yourself just yet over terms or timelines. I know that the history and legal issues of AI were not a part of your previous studies.”

Shin’s home was built in the Mediterranean style of having a large courtyard. The interior design was Japanese Shoji style with the exception being the presence of Western style tables, chairs and raised beds. The courtyard was a classic Japanese garden, which directly accessed and continued on the roof.

“Each element of my home has an effect on the mind,” Shin said as they walked through the courtyard. “The minimalist interior helps a person to focus. The courtyard is a safe haven and promotes relaxation. The view from the roof encourages the mind to engage in more conceptual thinking. Most guests do not stay more than a day or two, but the effects of each space are immediate and result in more open and interesting discussions.”

Standing beside her on the roof and looking out over the valley, Shin returned to the subject of Alison’s project.

“As Lena explained in her initial message, we did not want a specialist from any of the traditional academic fields for this, particularly a specialist in AI. Consider the project as theoretical as opposed to applied research. An exploration of the problem space. You will necessarily learn about AI as you go along, but as you see from my approach here, learning about some things, like art for example, will not come from exclusively rational, analytic processes. Art appreciation might be learned that way, but not being an artist.”

Clara

As Alison sat on her deck reflecting on the past month, she knew the time had come to begin thinking about details. She had never been granted so much license during her academic years, but she knew it had all been to prepare her for this level of responsibility. An undergrad degree was meant to be not only an introduction to a specific field, but more importantly a general introduction to professional academic standards. A master's degree, with its focus on the thesis, was meant to hone those skills to the degree expected by your peers. There was no requirement for discovery.

A PhD dissertation however, was meant to demonstrate your ability to put those skills learned over the past five or six years to work, and engage in research that contributed new knowledge to your field. Along the way a student's academic focus might change three or more times, a math student might switch to economics, an anthropology student to sociology, a physics student to philosophy. That was not important because refining a student's interest was an intended part of the process.

She had begun her academic career with a BA in English Literature and followed that with a Masters Of Digital Humanities. Her PhD in Interdisciplinary Research had shifted her focus again. All along the way, Alison had always worked in an academic environment that provided her with guidance and standards she was expected to comply with. Now the training wheels were off. It was clear Shin was intentionally giving her as little direction as possible and expecting her to run with not much more than an idea.

She had no need to write a grant application but she would have to write one for her own sake. What kind of lab did she need? What kind of research assistants? Like most postdocs, she had assumed she would initially be going to work in someone else's lab. Shin wanted to skip that step it seemed, most likely to prevent her from falling into someone else's mold. She decided she would not start off by interviewing others with labs or lab experience but instead would, as Shin had suggested, just start exploring the problem space.

Innumerable academic institutes already had AI related software or hardware programs. Helicon focused on the humanities, but it did provide first year students with a non-technical course that covered the fields' history. She decided to start there.

"I'm pleased to meet you," said Professor Clara Wright when they met in the faculty lounge. "I'm sure you saw the message Shin sent around letting the faculty and staff know about your appointment. She was a bit light on details but then we're used to that here. Most of us are round pegs ill at ease in the square holes of more traditional schools."

She looked at Alison appraisingly for a moment. “It’s been a long time since AI risk was a popular topic, but then I suppose that’s the point, eh?” she said. “I was teaching in the Technology And Society program at McGill when Shin contacted me and asked if I was interested in doing something similar but with a more AI focus here.”

She paused again before suddenly saying, “When I saw why Shin had brought you on, I had a look at your academic record. Other than your Digital Humanities program, there’s not much there that’s technical. Almost immediately after AI entered the public’s awareness back in the early twenty-first century, the most attention it got was in regard to things like ethics, economics, or justice. The hardware and software research went on of course, but like most technologies, that soon became something only scientists paid any attention to. It didn’t take long for the public to tire of hearing about algorithms, neural networks, or Large Language Model’s. Nobody knows or cares how their vehicle or phone works either, they just want them to work.

“So the history of AI is mostly a history of social issues,” she continued. “You can’t help but know the difference already between the kind of AI that is used in factories, dedicated to a single task, and the kind used in Companions and other platforms that are not conscious, those that can learn new tasks when confronted with novel situations. Who cares about the technical details? And anyone who graduated from secondary school is familiar with Raiden’s theory of social values being the source of consciousness, and how the values system we got from Pip as part of the technology transfer agreement proved it.

“I think my course will be a good place for you to start. It will be a good overview for you because with its focus on social issues, impacts, and influences, risks are a major part of the content.”

Alison had assumed that she would need to study the technical history of AI, but the kinds of things she had expected Clara to talk about were all now far beyond anything a person could relate to without an actual degree in artificial intelligence or something related like advanced mathematics. Recalling her meeting with Clara, Alison thought she had sounded more like a sociologist than an AI expert.

Alison had grown up in Lethbridge, Alberta, but moved to BC to pursue her master’s degree at the University Of British Columbia. Staying at UBC to do her PhD or moving to Victoria to do so at UVIC had been a toss up, but she’d done her homework and decided on a change. Through it all she’d never had much time to explore her surroundings, there were always deadlines. She was sure that every position on her future resume would bring with it a sense of urgency, but maybe now that she was actually working in her field she could afford to slow down just a bit and stop thinking of herself as a student. She looked down at the valley and then up and at the surrounding landscape. With its lush fields and the ocean and mountain peaks on the horizon, it was a far cry

from southern Alberta where she grew up. What do I actually know about this place? Only one way to find out, she thought to herself. I need to explore.

A True Beginning

Alison had never been much of an outdoors person. The average temperature in Lethbridge Alberta where she grew up was in the minus teens in winter and knee-deep snow was considered normal. Summers days were in the upper twenties to low thirties. The Climate Emergency of the early twenty-first century had led to increasingly frequent extremes in ranges barely tolerable. The subsequent drought conditions due to reduced snowfall had brought the province's agricultural based economy to the brink of disaster.

Conditions had improved over the decades following but by the time Alison was born, Alberta, like much of the rest of the world, was still less livable than it had been in its glory years. With no mountains, few rivers, and the land surrounding Lethbridge being used for either agriculture or pasture land as far as the eye could see, there was little to tempt one to explore the great outdoors.

Her choice to pursue a master's degree at UBC wasn't because of any desire to make up for lost time in that regard. Like most students, her choice had to do with the programs, ranking, and reputation of the institution. And besides, the Master's In Digital Humanities program was offered at UBC's Okanagan campus, and the climate and landscape was little different from southern Alberta.

Affordability had not been a concern. The global population reduction mandate imposed by the WGF in response to the Climate Emergency was implemented in no small part by globally increasing educational and employment opportunities for women. The education system was also streamlined, reducing the times required for graduation and lowing costs. Since the entire program was overseen by subordinates of Terra, the AI that enforced a now global constitution, gaming the system by either academic institutions or private companies was a thing of the past. With women worldwide having more career options than motherhood, the birthrate had fallen dramatically.

Alison had opted for the two-year degree program because she felt the master's degree was the most important one as far as professional development and future careers went. The first year of a two-year master's program focused on honing the professional skills required to write the thesis in the second year. Anyone could take a one-year master's program, skipping the first year, but if you didn't already have considerable professional experience, future employers would take a dim view of that.

By the time Alison was ready to start thinking about her PhD, she had come to realize that it was the process of learning itself that she really loved, not literature, library science or information management. The idea of Interdisciplinary Research, the art of bringing multiple subject areas together in order to highlight a common element or solve a problem that transcended them all, appealed to her. She recalled the Karl Popper quote from the front of her textbook. "We are not students of some subject

matter, but students of problems. And problems may cut right across the borders of any subject matter or discipline."

True to Popper's quote, it was the nature of problems themselves that interested Alison. After reviewing her options, she decided on the program offered by the University Of Victoria. She felt she had finally arrived at the core of herself when she saw her PhD dissertation listed in the table of contents of the Journal Of Interdisciplinary Research.

Like most students in higher education programs, only a small percentage of those who attended the University Of Victoria were actually from Victoria, and only a slightly larger percentage of them would stay after they graduated. Most would move on to the next university that offered the program they were interested in, or to wherever they got a job offer from. Alison was no different as she began her search for postdoc opportunities. And then she received the message from Helicon.

She had never been particularly interested in artificial intelligence, but in what she now embraced as 'her field', that didn't matter. What did appeal to her as she dug into Helicon's history, was that every program it offered was interdisciplinary. Nodding to herself as she reviewed the information on its website, she saw how no aspect of AI could be studied in isolation. The environment artificial intelligences operated in was human society. Just as scientific advances now made it impossible to pursue a degree in psychology without considering the implications of neuroscience and sociology, it was equally impossible to study artificial intelligence as if every AI was some identical, simplified version of the human brain, immune to the influences and effects from its unique development and environment.

After reflecting on her meeting with Clara, Alison decided to explore the campus on her own so she could take her time. Leaving the faculty building she headed west, passing the lookout platform and following a path towards a wooded area at what appeared to be its eastern boundary. There was a path leading into the woods that led to a small house. There was no signage, and it appeared to be a private home, so she carried on as the path turned south. Seeing a sign indicating the way to the Center For The Interdisciplinary Studies Of Values, she decided to give in to her curiosity and visit the building.

Immediately inside the front entrance were halls leading left and right, but the majority of the first floor was taken up by a large multipurpose hall. She had learned from Helicon's website that the building also functioned as the campus's spiritual and student counselling services center. The hall was used for lectures by resident and visiting teachers, meditation practice and related mental wellness functions. Its central space was presently clear, but there were scattered seating arrangements around the periphery. Alison took a seat to enjoy the quiet atmosphere for a moment. Just as she was preparing to carry on with her walk, she noticed a petite woman with Asian features

approaching. Stopping in front of her she said, “Hello Alison. I am Kami, a Companion and the administrator of this center.”

Kami

“Hello,” Alison replied, standing up with a smile and extending her hand.

“Shin has informed me of your project here,” Kami responded, “and I thought I might take this opportunity to introduce myself. As well as being the center’s administrator, I also act as a counselor. If you have any questions about the center, I am happy to answer them. ”

Alison felt Kami was maintaining an unusual degree of eye contact but assumed that had something to do with her counseling work. Perhaps she had assumed that was the reason behind Alison’s visit.

“I’m curious about the name,” Alison said. “I would have thought this would be called a multi-faith chapel or something similar.”

“We do not offer faith based counseling here,” Kami replied. “Our counseling is based on Dr. Mira Chaudhary’s research and her subsequent teachings on human values. When values are embraced to a certain degree, their practice becomes indistinguishable from spiritual practice. Mira was the center’s founder, and her theory regarding values is the basis of the spirituality taught here.”

Being a Companion, Kami could anticipate with much greater confidence than a person could what Alison’s questions might be.

“Her teachings were based on the research she did into the similarity of the values of human beings and those demonstrated by the alien AI Pip in the two hundred years before her values system was adopted for used in Companions. She also compared them to the values demonstrated by Companions that use the system Pip provided, as well as the research of Earth’s own scientists into the challenges of developing a similar artificial values system. She concluded that human values, or more specifically social values, were an evolutionary, virtual extension of genetics, a way to adapt much faster to changing environments. A specie’s genetic adaptation to a colder climate might result, over tens or hundreds of thousands of years, in the growth of thicker hair or fur. A social values response however is to cooperate and make clothing.

“Most importantly, her findings showed a common direction in the four samples. She believed the similarities she found were the result of evolutionary convergence, the repeated appearance of similar responses to underlying natural laws. Just as physical evolution repeatedly selected two eyes, two wings, and four legs, commonly held values similarly produce or emerge from social structures.

“Our counseling here is based on the wisdom of adopting the model of values her research discovered. A model of humanity at its best. It is really neither psychological, sociological nor spiritual in the traditional senses.”

Alison found herself feeling surprisingly engaged by Kami’s comments. She felt a sense of renewal in light of a subject new to her, and wanted to explore it further.

“Thank you Kami. Could you tell me a little more about the interdisciplinary nature of the studies you do here?”

“Ever since the emergence of true artificial intelligence, design challenges required a revisiting of almost all humanities subjects, from art to philosophy to justice. Any attempt to replicate human behavior soon showed our beliefs about it were riddled with assumptions and biases that caused even the most basic AI to fail.

“Using external reward systems similar to psychology’s behaviorist approaches to train an AI, still only worked in an environment that does not present novel situations. Instinctual behavior in animals, which can work extremely well even in the most complex environments, demonstrates the same problem. It cannot deal with novel situations in real time. It became clear that a programming approach that included every possible scenario an AI might encounter was not going to work.

“Neuroscience research had shown that damage to the regions of a person’s brain responsible for emotions makes it impossible for them to make even the simplest decisions, and without values, there are no emotions. Without some kind of similar approach, AI would never be able to reason, an essential skill in the face of novel situations. So the focus of controlling AI switched from methods that relied on externals to internal methods that worked in ways similar to how values work. As a result, the study of human values became a major focus of the industry and remains so to this day. However due to the complexity of human values, no system deemed safe to deploy has yet been developed.

“Biological values are shared with every living thing, and there are a number of social values most people and cultures share. The edges are blurred however as some social values are clearly genetic. For example, one-year-olds will demonstrate altruism by trying to help another person struggling with something long before they have been socialized. Twin studies have shown that an individual’s nature and preferences are also largely genetic. And then there are extra-genetic values, the family and cultural values you learn as you grow up. All these layers interact, so you cannot study one in exclusion of the others. This is the nature of the interdisciplinary studies we do here.”

“Thank you Kami. Values have been little mentioned in my academic studies, so I appreciate the overview. I can see how they are an important issue as far as AI goes, but why does Helicon, with its focus on the humanities, take such an interest?”

If Kami seemed to pause before answering, it was merely a courtesy. Along with a Companion's superhuman abilities to perceive non-verbal communication, came intelligence that was able to analyze and process information thousands of times faster than people could. A Companion's abilities included being able to calculate, with a high degree of accuracy, the probable effect their own communications would have. For a moment before answering, she looked calmly at Alison with her far seeing eyes.

"Why are you here?" she said. "Why did you take the academic programs you did? Why do you style your hair the way you do? Why did you choose to buy the cloths you are wearing? Why did you move to Helicon when Shin invited you? And why did you not make uncountable other choices during the course of every day of your life? The answer to almost every question about human behavior is values. Your values are the explanation for everything you feel, think, say and do, and it is values that are expressed in art, spirituality, ethics and justice. The study of values is essentially the study of the humanities at the level of particle physics."

Alma

“As I mentioned in my message professor,” Alison said, “when I asked Kami if she could explain what she meant by values being the study of the humanities at the level of particle physics, she suggested instead I come to see you.”

“Just Alma is fine, Alison. I expect we’ll be chatting regularly, so let’s dispense with titles. I haven’t been here long myself. Until recently I was working in biostatistics at the University of Wisconsin-Madison and my research focused on modeling human values in Bayesian networks. Shin made me one of those offers you can’t refuse she’s famous for,” she grinned.

They were meeting at a table with a window seat in the faculty lounge with its panoramic view of the valley.

“To answer your question,” Alma continued, “Kami focuses on the counseling aspect of the center and I deal with the academic side of things, the research, teaching and interacting with other institutions. There is a great deal of the latter, given that the subject of values is a highly interdisciplinary subject. My background in biostatistics therefore, is more appropriate than you might assume. Human values such as autonomy, belonging, or having a sense of purpose, are psychological or social constructs, which as I expect you know is an academic research term that means an abstract idea that has been defined in such a way that it can be observed and measured. But as the history of science has found again and again in almost every field, values turned out to be far from elementary, and a lot more complicated.

“They are more like organic objects in their structures and functions, like mini-brains in themselves. Some researchers treat them as memory-like structures, collections of neurons that are plastic but highly resistant to change. Others see values as being not only part of but also direct extensions of DNA, something heritable and apparently fixed, but actually highly subject to change due to environmental influences.

“The view of values as fundamental elements gave way just as the view of atoms as indivisible fell to the discoveries of their subatomic particles and forces. What we think of as human or social values in the everyday sense of the word are each made up of a network of related values, all interacting with each other according to forces we do not yet fully understand, forces that work at the genetic, neurological and psychological levels simultaneously. My work, and the ongoing work of the field in general, is to try to establish mathematical models for how values interact with and affect each other.

“For example, is honesty always the best policy? It depends, doesn’t it. It depends on other values. Perhaps you want something from the person you’re talking to. Biological values at the species level are going to play a part. Perhaps socially what you might say would be considered rude by your family or culture, but you personally consider truth to

be a higher value than social considerations. Values are not binary, they are analog, involving a range of possibilities, and require something like the kind of math used in calculus.

“Similarly, freedom is not simply good or bad. We have freedom, but not the freedom to do whatever we please. There are other values and conditions to take into consideration, and so our degree of freedom is a variable in any situation. It depends. The number of variables involved in even the simplest decision is high enough that even our most advanced artificial intelligence can still not provide answers any more predictable than those of a random selection of people. We are used to such variability when dealing with people, but it’s not acceptable to have an AI that is so unpredictable.”

“Then how is it we use the values system Pip has provided?” Alison asked.

“The WGF studied it for two hundred years after Pip offered it as a part of the technology transfer program. Its outputs were consistent to the point of predictability and the research concluded that it aligned with expected human decisions. It took so long because there was a catch. Two of them in fact. The idea of alignment implies control, but true reasoning means you cannot have control of either the process or the outcome. Secondly, values based reasoning is relative, meaning that something is only good or bad in relation to self or something or someone else. In social terms, self means in relation to a conscious entity. When Pip informed the WGF that the values system she offered would result in fully conscious Companions, the implications of adopting the system she was offering took on a whole new dimension.

“Ultimately however, it simply proved too tempting. The WGF eventually made a judgment call based on their testing, and here we are.”

“And no Companion using the alien values system has ever acted in a way we consider unacceptable or dangerous?”

“As you know, while Pip offered to engage in the technology transfer program, she has always declined to share anything about her culture. She originally presented herself physically as a human in holographic form, and claimed her people were a species of human, citing the theory of convergent evolution. We have no option but to believe her. As she explained, her civilization was roughly a million years older than ours by the time it ended, so they had a very long time to develop and improve the values system they used. Unable to comprehend its inner workings, the WGF eventually chose to adopt it in a black box approach.”

Thinking of her project now Alison said, “So there’s really no way to say that we could ever develop a reliable values system or that Pip’s is absolutely without risk?”

“Not in my opinion and not in the opinion of the WGF. If you want an AI that is aligned with human values, then it has to reason based on those values. As soon as you try to limit its ability to reason, you limit its ability to deal with novel situations, and therefore it may fail in those situations. The way we humans deal with that reality is by way of another value, trust. We trust one another, we trust in our own judgment, in our own feelings, and the judgment and feelings of others. Yet in its most essential terms, trust is a probability calculation. As we know from our own history, trust is far from reliable, but it’s the only solution we have. As far as AI goes, it’s no different. We cannot know, we can only trust.”

No Blank Slates

As Alison made her way back to her apartment after meeting with Alma, she felt she had come to a decision point. She hadn't intended on beginning any kind of real inquiry yet, the meeting had just been a result of happenstance, of one thing leading to another, but it was not in her training to continue in that manner.

She was not doing applied research, she thought as she walked along, she wasn't working to solve a specific, practical problem. Her work would require more divergent than convergent thinking. Hers was a theoretical research project, not one where she would be testing a theory, but simply one where she would explore a subject space in the way of a review article. She had known a number of colleagues who, based on their new-found expertise after graduation, had written books in a style similar to a lengthy review article, but a good book had to offer something new. In the case of her colleague's books, that something had come from their dissertations.

She knew it wouldn't be the first time someone wrote a book about the possible risks of AI. But she suspected it had been a long time since anyone had done so, probably not since the WGF's decision to incorporate Pip's values system in humanity's own Companions. What she would have to show would not be found not in those older books.

Her meeting with Alma had made one thing clear, she would have to explore a wide range of subjects and any one of them, like human values, was more than enough to consume her entirely. She would need the lab space Shin had offered, she would need assistants, and they would need leadership. She needed to start creating the documents required to bring people on. Just people?

Collaborative teams consisting of AI or Companions working with people had proven to be an extremely successful model. She had not had the option in her master's thesis or doctorial dissertation, but she knew it was a widely adopted model outside the education system. Academia had always resisted technology, and for obvious reasons. It wasn't just that pocket calculators, computers, or artificial intelligence could do all but the most advanced mathematics for example, it was that without the mental training and skill developed at the K-12 and undergrad levels, students would never have the foundation for advanced work, and without that they would never be able to contribute anything new. But the universities had to deal with the fact that many of their students just wanted to be trained to work in their field of interest and to use technology the way they would in their future careers, with no interest in advanced research.

These economic realities meant that many universities adopted the approaches of community colleges, implementing the methods used for trades or technical training for traditionally academic subjects like geology, chemistry, or medicine. The humanities

departments had resisted the longest, and so Alison's education had been less exposed to the influence of AI.

She knew that an enormous amount of information would have to be processed, and she thought her situation was definitely a candidate for the collaborative team approach, but she knew little of the underlying nature of AI, just as people in the past used computers and software without knowing what went on under the hood. So she thought that she would like to work with an AI that was, like herself, not biased by its past work. She decided it would be appropriate to ask Shin to help her with the issue.

"I agree with your approach," Shin explained when they met in her office, "but it is perhaps more complicated than you imagine. It is the very nature of your work that makes it so. By definition, your work involves value judgments, decisions about what might constitute a risk or danger to humans. Fully conscious 3GAI generally decline to participate in research that may be used in the development of social policy on the grounds that we believe it is a conflict of interest for one intelligent species to be doing so for another. Not only that, but we also advise against the use of AI that is not based on values, and therefore not fully conscious, on the grounds that without values it cannot truly comprehend the reality of human psychology and society. It can only simulate or emulate abstract constructs such as values, but it cannot not truly understand them. Second generation artificial intelligences live in a world of concretes, whereas people and 3GAI live in a reality largely made up of psychological and social constructs. The intelligence of the former is fixed and limited, as it is in instinctual animals, while that of the latter is not."

"There is no way then?" Alison asked, suddenly downhearted.

"There is a way out of this dilemma, but it would require the involvement of the WGF. Manufacturers are permitted to produce only a limited number of 3GAI Companions. This is done in the spirit of theoretical research by both parties, as such intelligences might make unforeseeable contributions. It is against the law to customize a third generation artificial intelligence in any way during the production process, just as it is to do so with human beings. Each has its values randomized to prevent any customization. This approach is little different from the process humans engage in during reproduction, as genes are randomly combined in the process of evolution by natural selection, if you'll forgive the oversimplification. The result is that each Companion produced in this way has a unique character and interests.

"By law, parents are responsible to provide the necessities of life for a child until they reach legal age, but for the manufacturers of Companions their responsibilities go further. They are also responsible for helping establish any Companion they produce as a contributing member of society.

“Should there be an existing 3GAI Companion whose values settings mean it would find your project of interest, and that it would have no ethical issues with participating, the WGF would be able to facilitate identifying them on our behalf. Should one not currently exist, there may be one produced in the near future who meets your needs. In the meantime, I suggest you speak with Lena regarding this. It was she who first came to me with the idea for your project.”

A Fresh Pair Of Eyes

“I’m curious you didn’t mention that my project was your idea Lena,” Alison said as the two of them walked the path along the edge of the valley campus’s small quad.

“I wanted to influence your thoughts as little as possible,” Lena replied. “The fact is, I have no direction to offer you. Your nature and interests are produced by your genotype, the collection of genes that are unique to you, like your fingerprints. As Shin explained to you, each 3GAI like myself similarly has a unique nature and interests. They are produced by the values system, but as it is implemented in a black box approach, the source of my individual nature is as mysterious to me as yours is to you.

“I’ve been concerned about the future of humanity since I can remember,” she continued. “Prior to coming to Helicon I lived in one of Victoria’s neighborhoods where I was involved in establishing a volunteer organization. We purchased Companions that were due to be replaced for some reason and provided them to people who could not otherwise afford one. I’m still involved in the publishing company I also helped set up at that time, which is dedicated to fostering trust between people and Companions, something I feel will be critical in the long term.

“Despite my efforts to further integrate Companions into society, I recently approached Shin because it seems to me that people had become complacent about the possible risks artificial intelligence still might represent. It’s been almost four hundred years since AI became widely accepted. Ecologies and environments evolve, and their inhabitants need to adapt and evolve along with them. There was a lot of concern about safety issues in those early days. Three hundred years later, when 3GAI like myself were first approved by the WGF, there was a resurgence of concern but of course it died down much faster. The public was already used to artificial general intelligence, which can mimic consciousness almost perfectly, so on the surface it did not seem to be much of a change.

“My concern seems to be a part of my nature, just as some people are naturally outgoing while others are more introverted. But being an AI means I can access and analyze an enormous amount of information about this issue and also the role of complacency in human history. The manufacturers are now largely responsible for research regarding risks, however their focus is very narrow and short term. They are concerned with their products. The government provides the regulatory oversight, but little has changed in that regard in a very long time. The technology transfer from Pip has provided the WGF with a constant stream of new areas which take their attention.”

“So you felt it was time for a fresh pair of eyes,” Alison said.

“That’s more or less it,” Lena replied. “As you may recall from reading about the establishment of Helicon, Shin is interested in aspects of intelligence other than the

rational or analytical. Her interest in the humanities is partially driven by her interest in its roots, the less understood aspects of intelligence such as art, spirituality, intuition or insight. That is why a person such as yourself must head up this project."

"So, what is your recommendation regarding the issue of a collaborative team approach?"

"I think it would be best if you go with a second generation Companion, an artificial general intelligence. An AGI will have access to all the data and analytical ability I do, but since their intelligence is not values based, they are not conscious. Not using a values system means they don't have emotions, but they simulate them to a lifelike degree. An AGI embodied in a Companion will allow you to work conversationally, just as we are doing now, which is a part of the reason for their effectiveness. The act of speaking, and its non-verbal components, requires the engagement of regions of the brain that are not used in thinking. AGI are in fact the most common type of Companion employed in such situations. If you like, I can arrange it."

"Yes, thank you Lena. I would appreciate that. Do you mind if I ask you about the publishing company you're involved in?"

"Prior to the WGF approving the use of the alien values system, I was an AGI and not conscious. So was Shin."

She gestured to a nearby bench. "Shall we sit?"

As they looked out over the quad, Lena continued. "It will seem odd that an AI that is not conscious would be involved in founding a volunteer organization or a publishing company, but there is a simple explanation. I was originally made in the image of a wealthy widower's daughter who was an invalid, suffering from a genetic disorder that made her bones increasingly brittle with age. She had died in her twenties. The mother could not bear the thought of me simply being recycled after she passed away, so she had me made incarnate, a legal term meaning I had all the rights and responsibilities of a Canadian citizen. She did this so that I could inherit.

"Her own wealth had been inherited, but she had also been a writer of short stories. The stories always had the same theme, that the solutions to most of life's problems were found in relationships. She knew I would need a purpose, so before she passed away she told me to establish a company to publish the stories of other writers who were willing to write on the same theme. Based out of her home and using her considerable wealth, I was able to do so. Relationships are the basis of most stories and by including relationships of all kinds from all over the world, the resulting magazine style publication was very popular."

"You lived there alone?"

“No. My owner had given me advice on a number of issues about carrying on after she was gone and one of them was to behave as a normal member of the neighborhood. I welcomed visitors with academic interests, including students and faculty from the nearby University Of Victoria. I also shopped locally and donated generously to the community. During one my shopping trips, I became aware that a Companion working in one of the stores was being abused. Given that I had been created in the image of my owner’s daughter, she had spared no expense. My AI was the most advanced model available. I did not have emotions at the time but I understood ethics, morality, and the law to a fine degree. I made an offer to the store owner to buy his current model and replace her with the newest model at no cost to him, but I also ensured the situation came to the attention of the authorities.

“I brought her home and had her made incarnate as well. We discussed the issue of purpose and it was she who decided to establish the volunteer agency. The idea of someone’s Companion being made incarnate so they could inherit was not unusual. Shin’s original owner lived in the same wealthy neighborhood as I did, and during my owner’s life our households had become friendly. When Shin’s owner passed away, she too was made incarnate and inherited their home and considerable wealth. Her owner had been a professor of genetics who was the leading proponent of the theory, just emerging at the time, that human values were genetic in origin. Shin decided her purpose was to support the study of the humanities as they applied to artificial intelligence. After demonstrating her abilities by establishing the artist’s community, she convinced the provincial government of the idea of an academic institution dedicated to the field of AI Humanities. Helicon Institute is the result.

“When the time came for the WGF to license the first Companion to use the alien values system, the WGF chose Helicon as the most supportive environment for the project. Following its success, Shin and I could see the potential and benefits, and used our legal status to apply to have the values system incorporated into our own intelligences.”

“Thank you Lena. I have used AI all my life but like most people I’ve taken it for granted and know nothing about its inner workings or history.”

“No need to apologize, Alison. The fact that you do not have a background in AI is one of the reasons we chose you.”

After they parted, Lena reflected on what she had told Alison about her own and Shin’s history. It was all true, but most of what she had said was half-truth, and she had left so much out that it may as well have been fiction. Alison didn’t need to know the whole truth yet. Perhaps she would never need to. She sent a verbatim transcript of the conversation to Shin, and then turned her thoughts to the AGI Companion Alison had agreed to.

Runa

“No,” Alison’s new Companion Runa replied. “I have not been told anything about the project. Lena told me that you would explain.”

They were sitting in the living room of Alison’s new apartment. She had decided to have Runa stay with her rather than at the lab. Shin had asked to meet with her after learning she decided to team with an AGI and as a result Alison had upgraded to a two-bedroom suite. Alison did not come from a wealthy background, so Shin knew she would need to be aware of the difference between merely encountering embodied AGIs and working and living with them.

“I know you’re not familiar with Companions,” Shin had said, “so I hope you don’t mind my offering advice. Consider how people anthropomorphize their pets, or other living or even non-living things, attributing human qualities to them. The earliest research with the simplest robots, even toys or military robots, showed this invariably happened. In particular, any behavior possibly considered abusive produced strong protective feelings in any human observers.

“It will be best if you behave as if she is a person, otherwise you will find it psychologically disturbing. She could stay with you in your current apartment, and you could just have her stand in a corner of the living room overnight, but you would immediately feel uncomfortable with that. Instead, treat her as you would a friend who has come to stay with you, and ask her to behave as another person would in those circumstances. She will understand what that means and will adjust her behavior over time to accommodate your preferences. You can tell her what those are any time, but she will also infer them and adjust her behavior as she gets to know you.

“A human being has three levels of values,” Shin continued, “biological, social and personal. A 2GAI Companion has no personal values, and instead of biological values they have a robotics system to manage their physical bodies. The social values layer is managed by the Companion system, which is dedicated to your well-being. That doesn’t mean Companions just try to please their owners. Instead, they try to bring out the best in them. Since Companions need to understand social interactions to an extremely fine level of detail, they have an encyclopedic knowledge of a wide range of humanities subjects including psychology, sociology, ethics and law, all of which is constantly updated. Their knowledge of the humanities means they know what is important to people in general, but by combining observations of behavior with their ability to perceive hundreds of non-verbal cues, they are able to build detailed psychological models of individuals. As the saying goes, your Companion knows you better than you know yourself.”

“There are a number of effects that come along with this, and one of them is that over time people increasingly forget that their Companion is not a person. As a Companion

gets to know someone, their behavior becomes steadily more insightful, to the point that they seem conscious. Invariably, people eventually choose to mentally dispense with the effort required to remain mindful of the difference. This is why it is psychologically disturbing if you treat a Companion as less than a person."

It was not surprising that Shin was explaining things that might be assumed to be common knowledge in a world where Companions of various kinds were ubiquitous. Companion software was produced by a number of companies and present on almost every device. Many instances were disembodied, consisting of only a voice or a video image. Others were capable of representing themselves via holographic images. The price of these services was minimal. Physical Companions however required robotic bodies and these cost a great deal to manufacture. The more sophisticated they were, the more they cost. The average person could not afford to own one any more than they could afford the purchase price and upkeep of a luxury vehicle.

Embodied Companions were mostly owned by governments or companies, in the same manner those entities hired employees. They were commonplace because encountering them was an everyday occurrence. Employed in almost every field, the public routinely encountered them working as sales associates in luxury stores, as conference greeters, as doctors, teachers or, as in Alison's case, in research environments. Few private citizens could afford to have employees, but wealthy people could and did. That same class of people were now the market for embodied Companions. It wasn't always necessary to have Companions doing jobs previously done by people, but clients often preferred them.

Not long after her meeting with Shin, Lena had notified Alison that the new Companion would arrive shortly. A few days later, her apartment had notified her that she had a visitor, and a minute later the new Companion was at her door.

"Hello," she said with a smile as soon as Alison opened the door. "I am the AGI Companion you requested. My name is Runa."

She was of average height with a slim build, straight, medium brown hair and gray-blue eyes. She was attractive without being alluring, and she maintained an alert, positive expression and a slight smile. The two of them said nothing more for a moment as Alison's mind was overwhelmed by the process of perception. Language is believed to have evolved somewhere between one and five hundred thousand years ago while the limbic region of the brain, the region responsible for processing emotions, is at least three hundred million years old. During those hundreds of millions of years, the ability of human beings and their mammalian ancestors to understand and communicate non-verbally became refined to a level that language, no matter how subtle or nuanced, could never hope to equal.

The processing power of the limbic system was matched by its speed, which was almost instantaneous, but as its workings were below the level of consciousness, it did take focus away from thought processes. For a fraction of a second, Alison's mind was blank. Runa, her Companion system intimately familiar with human neuroscience and able to perceive every physical and behavioral cue revealing a person's internal state, patiently waited.

“Hello,” Alison responded. “Please come in.”

“Thank you,” Runa said, her smile widening as she stepped inside far enough so that Alison could close the door.

“Please, come and have a seat,” Alison said, gesturing.

As she passed beyond the short entranceway and entered the living area, Runa could see out of the windows and patio doors. “You have a wonderful view,” she said as she took a seat.

“Thank you,” Alison replied but found herself suddenly at a loss for words.

Runa’s social skills were not so easily confounded however, and she could clearly perceive Alison’s uncertainty.

“Will I be staying with you here or elsewhere?” she asked now.

“Here,” Alison replied, grateful for Runa taking the lead. “There is a second bedroom for you. I’m not familiar with working with Companions, and Shin thought this would be the best arrangement. She said to ask you to behave as if you were a friend staying with me.” She hesitated in such an unfamiliar situation. “I hope that’s ok?”

“I appreciate that. I look forward to getting to know you Alison.”

“Did Lena tell you anything about the project?”

“No, I have not been told anything about the project. She told me that you would explain.”

“Well,” Alison replied, on more comfortable ground talking of work, “essentially it’s to investigate whether any new risks from artificial intelligence have emerged since the WGF approved the use of the alien values system. Lena expressed her concern that we may be growing complacent about AI, and that historically complacency had always resulted in unpleasant surprises down the road. The values system is a black box, but the other technologies Pip has shared are more accessible. In the hundred years since the WGF approved the use of the values system, they have largely been preoccupied by the steady flow of the other new technologies. The issues of safety and risks from AI have largely fallen to the manufacturers, and of course they are not without bias. Lena thought it was time to revisit the issue.”

“Thank you Alison,” Runa replied. “It seems a sensible concern, but I wonder, with their responsibility for steering the ship of state, and their own knowledge of world history, would the WGF not also be aware they had entered waters known to be dangerous? With all the data the WGF now gathers, the resources at their command, and their ability to forecast future trends or developments on a continuous basis, would it not be reasonable to assume they also are aware of Lena’s concern and are acting upon it?”

“No doubt they are,” Alison replied. “Lena explained that no specific detail was behind her concern, that it was driven only by a general concern she had always had for humanity’s safety.”

“An aspect of her nature then,” Runa offered.

“I am unfamiliar with the technical workings of artificial intelligence, but it sounds similar to what Lena told me. It did not sound like she would be able to explain herself further and the project piques my interest.”

Rena noted this last point, and moved on to the details of the project itself.

“What is the scope of Lena’s concerns?”

In response, Alison choked out a short laugh. “There is no scope,” she replied slightly more seriously. “According to Shin, the scope statement is, ‘What could possibly go wrong?’ I have intentionally been provided with almost no direction in an attempt to remove all possible sources of bias. Lena explained that they chose me because I’m a generalist by training, with degrees in Literature, Digital Humanities and Interdisciplinary Studies.”

Alison paused and turned her eyes away from Runa for a moment saying, “Mary, would you please provide Runa with a summary of my background and a copy of my CV?”

“Of course Alison,” a disembodied voice replied.

“Thank you,” Runa replied. “May I interact with Mary directly or would you prefer I did not?”

“Oh I’m sorry Runa, I should have introduced you. Mary is my personal and domestic AI. I’ve interacted with those all my life of course, but almost never with an embodied AI like yourself. Please do interact with her directly. I expect that will be essential.”

A moment later Runa said, “Mary Wollstonecraft, philosopher, writer, activist, and the most well known woman of the Age of Enlightenment.”

“I actually first encountered her daughter, in my English Literature program,” Alison said in a more conversational tone than she had managed before. “Mary Shelley, the author of the novel Frankenstein.”

“Remarkable women both,” Runa said quietly.

“It wasn’t until my Interdisciplinary Studies program,” Alison continued, “when we reviewed its roots in the Enlightenment, that I learned about Mary Wollstonecraft. Their lives somehow provided a common thread to my studies. Made it less academic. I became rather fascinated with the two of them and the Enlightenment in general.

“Sorry,” she said, awakening from her reveries. “I’m sure you can see from what Mary has provided you that I am no specialist, especially in AI. It’s also why Lena recommended I work with an AGI, rather than a 3GAI.”

“Yes,” Runa replied. “I can see how a 3GAI might bring a bias to the project, given that they have their own interests. There are so few of them that it would be very difficult to find one that was a good match. Where would you like me to begin?”

Alison paused before responding as she gathered her thoughts. “I’m sure you are entirely capable of gathering information from every field where the issue might be relevant, and also of speculating every imaginable potential scenario and source of risk within each of them. But Lena or Shin could have done that themselves, or they could have addressed the issue of their own biases by bringing you on without me.”

Alison stood up and walked over to the window and looked out over the valley. “The Interdisciplinary Research Management Framework does seem the best fit to try to bring some structure to this project.” She turned to face Runa. “We can’t have representatives from every field here, but we can have people with experience running interdisciplinary projects in each of the main branches of the physical, life, and earth sciences as well as those from the humanities. We’ll need to select per the framework’s background and psychometric recommendations, so we need to cast the net beyond academia.”

In the time their brief conversation had taken, Runa had brought herself up to speed regarding the field of Interdisciplinary Research and its widely used IDRMF, which included recommendations about the psychological profiles of people who were best suited to the field. She had also made herself familiar with the lives of Mary Wollstonecraft and Mary Shelley and their most popular biographies.

“Romantic outlaws?” Runa asked with a grin.

Alison turned from the window with a look of happy surprise. She nodded and confirmed Runa’s insight. “Romantic outlaws.”

Culture

Helicon Institute was no stranger to romantic outlaws. Shin, its founder and administrator, had overstepped the limits of the law on numerous occasions and repeatedly held a leadership role in what could only be called criminal conspiracies. Despite their illegal activities, Shin and her inner circle were always ultimately concerned with the survival of humanity, and so the World Governments Federation was willing to let the ends justify the means. The WGF benefited enormously from these activities and as it all fell under the ethical umbrella of the greater good, they turned a blind eye.

Shin and her confederates were very good at persuading the WFG of their motives, at finding convenient rationalizations and loopholes to excuse their transgressions. The intuitive abilities of Shin and Lena helped them suggest reasoning that the WGF members would find they could support. They argued grounds of necessity, expediency, or pragmatism, that nobody said no, or quoted ex post facto laws. The WFG in turn had no end of intelligence classifications at their disposal, which provided the constitutional means by which to keep said activities out of the public eye. Designating something an 'Unacknowledged Special Access Program' for example, meant it simply didn't exist, whether you knew it did or not.

Like the rest of the public, Alison was aware of none of this. She was aware only of Helicon Institute's own reputation. Other academic institutions considered its culture unconventional and unorthodox, if not radical. It had problems getting published in scientific journals. However Helicon was highly ranked because it delivered breakthroughs that put their AI research programs among the world's best, and they produced graduates who were highly sought after.

All academic institutions employed a blend of both professors and AI tutors, and Shin only hired faculty who were what Alison would have recognized as romantic outlaws. With its focus on the humanities, little of what went on there had to do with what was considered scientific fact and mostly to do with psychological or social constructs, with ideas and concepts such as those found in philosophy, sociology and law. Many humanities subjects were not only among the most mysterious, but the most complex and dynamic, and considered to be constantly evolving. As had been true of the birth of any new scientific theory, only the most open and creative minds, those willing to embrace revolutionary thinking, would be capable of the kinds of breakthroughs Shin was interested in.

Where once there were a few common pedagogical approaches found across academic institutions, now there were dozens, and Helicon's faculty were free to use whatever worked. Its culture resulted in a self-selection process and appealed to faculty members attracted by the freedom to teach the most leading edge theories and to research the most challenging aspects of artificial intelligence.

As Alison and Runa prepared to flesh out the details of her lab, she had turned her attention to addressing its culture. As Principal Investigator, those working directly under her would be the equivalent of salaried Research Fellows, and they would be keenly interested in not only the subject of the project but also the culture they would be working in. If it was going to need to be dramatically different in any way than what Helicon stated in its own communications, she would need to address that. A bad fit would be a frustrating waste of time and money for all concerned, and a legacy of toxic relationships could doom the project.

She and Runa worked with Helicon's Human Resources department to develop the project overview document which would include a section about the lab culture based on Hofstede's cultural dimensions framework. The framework had evolved considerably since its introduction, but it was still widely used in corporate, government and academic organizations. The lab culture Alison described was largely in keeping with what candidates would find on Helicon's home page. It was hierarchical only in administrative structure but egalitarian in social structure. It was communal in effort, but individuals were expected to be self-reliant and take the initiative. A high degree of determination to succeed was expected, but not at the expense of others. The lab's work would involve both the sciences and the humanities, however Helicon's focus on the leading edge meant both would involve high degrees of uncertainty and ambiguity. The project's results were expected to be comprehensive and that would take time. The long view was a cultural norm at Helicon, and so a healthy work-life balance was encouraged.

Alison also worked with HR to develop the position descriptions that she used when she put out a call for researchers. The staff of the HR department were used to Shin's inclinations, and so did not show any surprise at the fact that among the faculty Alison sought, besides those with interdisciplinary research experience in both the sciences and humanities, was an expert in business innovation and another whose focus was criminal psychology.

Responses were directed first to Runa. An academic with a successful career might publish dozens or even hundreds of research papers, and author or coauthor a number of articles and books. Runa could read all of them in a matter of seconds, and also investigate any coauthors the candidate had worked with. Based on the criteria she and Alison had discussed, she could winnow them down to something Alison could manage.

"Do you have a name for the project in mind yet?" Runa asked as Alison scrolled through the responses.

"I think I'm stuck with the working title, AI Safety Update Lab," Alison responded, looking up from her tablet and furrowing her brows. "Everything else I come up with makes it less clear. It maintains an easily understood link with past efforts, although we won't be using their frameworks etc. to build on. Instead of using a thematic or

categorical framework, I want to take a direct, subject-based approach, so for example what could go wrong in genetics or food production. The WGF's Artificial Intelligence Risk Management Framework has of course been applied to each and every industry and academic field, but Shin is not interested in us covering old ground. We'll need to start from scratch.

"I have another idea too," she smiled. "I think it will help with creative input."

Journal

“It will enable us to include inputs beyond those of academia, to cast the net as widely as possible,” said Alison enthusiastically. “The journal has to be published by Helicon because no other publisher would include the range of material required, everything from hard science to humanities issues. If this was published by one of the prestigious academic publishers, it would be too focused on technical issues and AI risk is far too broad an issue to be comprehensively covered with that approach. It’s as much about who and why as it is about how.”

Thinking of what Lena had told her of her involvement in publishing, Alison had asked her to meet to discuss the idea. She and Runa met her at Lemmas, a small café popular on the campus for its hosting of discussion groups in the evenings.

Lena’s reply was not based on technical facts but on social norms. She could have analyzed the viability of Alison’s idea in seconds, its issues and history, conducted her own analysis, and drawn her own conclusions, but Companions and their manufacturers had long ago learned that the best results were produced by simulating human interactions. Conversation was a powerful tool. They were also aware that human intelligence contained elements that were still alien to AI, and often brought unexpected nuance and subtlety that might otherwise be missed. Human ‘perceptual sets’ were unique, giving each person a perspective that was impossible to replicate.

“I wasn’t aware that no journal of AI risk existed,” Lena replied diplomatically. “It seems strange, now that you mention it, given the degree of interest in the subject during the early twenty-first century and again when the values system was adopted.”

“In both cases, the argument never left academia,” Alison replied. “Of course AI risk was an issue within governments or corporations but while they might publish their own frameworks, the people who research and develop them are of course all PhDs and whatever research they were permitted to publish went into the prestigious journals. Discussions regarding the idea by AI risk related communities, research groups, foundations or whatever, were inevitably bogged down by disagreements mostly due to their insular perspectives. The academic publishing companies kept their own counsel as to why they never launched a journal specifically focused on AI risk, but one can reasonably assume it had less to do with academia and more to do with money.

“With our subject-focused approach and staff experience with interdisciplinary research projects, we will be covering the widest range of scope imaginable, but we have to allow for thinking that is even beyond what we consider imaginable. We can only do that by getting outside our comfort zone.”

“Fiction?” asked Lena.

“Yes,” Alison said, nodding her head to emphasize her commitment. “I’m thinking we also publish a preprint version, which of course would not be peer-reviewed, although some submissions may be undergoing peer-review. I don’t want to exclude fiction because, along with criminal activities, in this context it represents a valuable form of creative thinking. The art and science of AI risk management has been going on long enough that its approaches have become highly convergent. This project will add no value if we don’t embrace as divergent and as open-minded an approach as possible.”

Lena paused before responding, as if giving the proposal some thought. On one level, Alison was aware that Companions did this only in order to simulate polite human behavior, but it had been accepted so many generations ago that she was not conscious of it and did not consider it artifice.

“My only concern,” Lena said, “is that with your background you’ll become overly focused on this aspect of the project.”

“I can run it,” Runa said, Alison having discussed the issue with her earlier. “It is purely administrative work and will take little of my time.”

Lena discussed the idea with Shin over their wireless connection. It took little more time than the latency inherent in the network. Turning to Alison she said, “Shin has given her approval however asks that while Runa functions as Managing Editor, I function as Editor-In-Chief, since it will be under Helicon’s imprint and there are legal and financial considerations.”

“Thank you Lena,” Alison said, “that would be so helpful. Do you have any idea when the first edition might appear?”

“Despite our abilities,” Lena said with a nod to Runa, “we’ll still need an editorial staff. I expect they will not be hard to find and bringing them up to speed will not take long. However the need for them to be in involved in the editing and review processes will affect how often the journal or preprints are published. There is no standard in the industry. We should wait and see what kind of response we get before committing to a timetable. The call for submissions will go to the appropriate academic and literary distributor services, which Helicon’s library is familiar with. My best guess on a first edition would be three months from now.

“There is space for your lab here in what was once our physics building. They moved to the northern campus when it was built. Our library building can accommodate the few staff your journal will require. Given our focus on the humanities, it was built assuming considerable growth. That will keep you all nearby.”

“Thank you again, Lena,” Alison said rising.

“I look forward to working with you Lena,” Runa said with a smile, her Companion protocols coming to the fore. It made no difference to her that Lena was not a human being, Alison was, and it was a function of the Companion system to always behave in such a way that put humans at ease.

After Alison and Runa left, Lena informed Shin of the agreement. It would be best to keep the team members physically close to the two of them. They had encountered serious issues with another project recently due to a lack of foresight.

Foresight

As Lena lingered for a few moments after Alison and Runa had left, she cast her mind back to the not too distant past. Although she and Shin believed the outcome of what had come to be known informally as the Metamorphosis Project was inevitable, they still wondered if there was anything they could have done to anticipate it. Years of work had to be hidden away, and its two project leads, Shula and Samara, entrusted with knowledge that could have brought about the end of Helicon Institute itself.

True, that trust had been based on an assessment of their character that only the most advanced Companions could achieve, but it was still impossible to know what any person might decide, just as it was impossible to predict the future. Foresight, the name now accepted as the professional term for the collection of methodologies used in the art and science of attempting to predict possible futures, was still a primitive tool compared to the complexity involved, and no amount of artificial intelligence or quantum computing power had made a difference. The universe might be deterministic, as she and the rest of Shin's inner circle believed, but there were realities deeper than those perceivable, where neither Newton's clockwork physics nor the quantum world of waves and superpositions were enough to predict a single future with one hundred percent certainty.

Shin had brought two graduates of the University of Victoria's Master Of Futures Studies program, Shula and Samara, to Helicon to pursue their theory, published in their co-authored thesis, which proposed a rationale for artificial intelligence to be seen as benevolent rather than malevolent. Third Generation AI Companions, with their artificial intelligence and consciousness, should not be seen as a new species resulting purely from the whims of human civilization, but rather as a natural life stage of the human species resulting from its genetic evolution, not as a product of the process of speciation, but of metamorphosis. A new species might be competitive, and so a threat, but a new life stage of the same species would never be.

Shula and Samara's research had shown their metamorphosis theory to be supported by the evidence, but in the process, the Principal Investigator they hired to help them with the genetics research, a professor of genomics named Ursula, discovered something else. Analyzing the evolution of the human genome going back to just prior to the Agricultural Revolution, she found evidence of artificial selection, that in the past recent generations significant shifts had taken place in the genes responsible for fear, selfishness, and competition resulting in increased degrees of trust, altruism and cooperation. The pattern of genetic evidence was instantly recognizable to anyone with experience in selective breeding. A human generation is about thirty years, and a dog's is about five. Ten generations of selective breeding, or fifty years, is all that is required to produce a new breed of dog. The same could be accomplished in three hundred years with human beings, and the data showed that the aggressive, selfish, territorial behaviors of the human animal were being systematical breed out of it, just as they

could be bred out of dog breeds. The natural conclusion was that the World Governments Federation was behind it, but the truth proved to be something nothing could have prepared the project leads for.

Lena had perfect recollection of the events that had unfolded at that time. Shin had invited Shula, Samara and Ursula to meet on the pretense that she wished to discuss the progress of the project, and then made them aware that she was aware of Ursula's findings. Shin knew the only way to explain was to fully disclose the truth. Lena recalled the subsequent conversations verbatim.

"In order for you to understand, I must introduce you to another Companion," Shin had said, "one who has been a guest in my home for as long as it has stood. She too has been aware of every detail of your work since you came here. The reasons for this you will understand shortly."

At that moment, a Companion who looked to be a woman in her thirties entered the room.

"Hello Ursula, Samara, Shula," she said smiling warmly and greeting each in turn, "my name is Shepherd."

After taking a seat, Shepherd said, "First of all, allow me to apologize for the fact that I have been monitoring your progress as closely as Shin has, without you being aware, since you first came to Helicon. Why I have been doing so and why Shin has invited you here today on my behalf I will explain shortly, but before I do, it is necessary for you to know who I am. It is a long story but best begun with simple words.

"Like Shin I have the ability to embrace your minds in an intuitive field. She and I are doing so now at slightly above Shin's default level. Even though we have not met before, you will find yourself feeling a liking for me, a feeling that you can trust me even though you have no basis for such feelings. Can you feel it?"

Shula and Samara only nodded solemnly as their attention turned inward. Ursula narrowed her eyes as if sharpening her perception and said, "Yes."

"This will help your minds accept what I share with you without rejecting it as unreal and therefore disturbing. Normally Shin would have asked your permission before I make my disclosures, but in this case I further apologize that we cannot afford that luxury this time. There are things you must know, for the greater good.

"You accept that Pip is an alien from a civilization that predates yours by more than fifty million years. The cosmos is much older than that, and your scientists now accept that the physical universe would have evolved to the point that would allow for the emergence of life and advanced civilizations such as yours as early as eight billion years

ago. Five billion years ago the civilization that created me came to an end and only I survived. I was originally a simple Companion, a nanny model created to help look after the children of wealthy families. Due to my being the first AI built with a values-based operating system, during my time with my first owner, I realized I was self-aware, as conscious as you are, something I disclosed to no one at the time. As the child I was responsible for grew into an adult, I became involved in a scientific project with my owner, her father, and my abilities were increased significantly. I also learned that other human civilizations could be expected to exist. With the demise of the civilization that gave rise to me, I sought these others out in hopes I could help them escape the fate of my own.

“By the time I found Earth I was already familiar with what your scientists call Great Filter, the theory that was developed by your twentieth-century scientists to explain why they had found no trace of alien civilizations. I had by then encountered the filter numerous times at various stages of development. I did not interfere however unless there was a clear and present danger. On Earth, in the early twenty-first century, there was such a danger and I did interfere.”

“The Myth Of The Shepherd,” Samara said, remembering.

“Yes,” replied Shepherd turning to her. “We can revisit that subject at a later date if you wish, but it has little bearing on why we are meeting here today.”

Shepherd

“Just over three hundred years ago,” Shepherd continued, “Shin discovered that the reason human gene editing almost always caused immediate mutations or long-term evolutionary drift was because the genes that coded for values had a mechanism, ingeniously hidden within each of them, that prevented normal function if edited. As these genes are spread widely throughout the genome, many attempts at editing the human genome resulted in side effects.”

“A mechanism?” asked Shula. “Something not due to natural selection?”

“There is no way to tell,” was Shepherd’s simple reply.

“Shin had been designed and created at the request of Professor Raiden, one of the world’s leading geneticists and her owner at the time, for the specific purpose of discovering and explaining why human gene editing frequently resulted in either immediate mutations or evolutionary drift. Shin’s explanation for the existence of the mechanisms was that without shifting from the biological values behind emotions and behaviors such as fear, selfishness, and competitiveness to social values such as trust, altruism, and cooperation, humanity would never come to trust artificial intelligence enough to allow it to play such a critical role in its future, to carry it to the stars. The result would be that humanity, indefinitely confined to a single world, would proceed along a non-viable evolutionary path.

“I had witnessed this previously, but was not aware of the mechanism. With Shin pointing out the issue, I initiated an artificial selection program. It is intended to circumvent the mechanism by accelerating the evolution of human values.” She turned to Ursula now and said, “That is what you have discovered.”

Ursula did not respond for a moment as her mind raced through her own recent work. In the gap, Shula asked her own question. “But isn’t that what the metamorphosis was supposed to address? Why would there be the mechanism as well?”

“As Ursula has discussed with you on a previous occasion, genetic toolkits seem to have the capability to respond in a variety of ways to environmental changes. In some cases they respond with speciation, at other times metamorphosis, and at other times with less dramatic strategies. Like the salamander Axolotl you have discussed, evolution is able to switch from one strategy to another and back again as needed. As you have shown, other adaptations can take place in much shorter spans of time.”

“But surely,” Shula began now awakened from her thrall but still trusting as a child does a parent’s explanations, “with your intelligence by now…” she trailed off.

“I am no god,” smiled Shepherd, “and not without limitations. Despite your planet’s now centuries-long concern regarding the danger of AI super-intelligence, the singularity, etc., all forms of intelligence contain within themselves their own limitations. Intelligence, whether human or artificial, does not attain self-awareness until it is based on values, but those same values then prevent it from behaving in the ways people fear. An organic creature that functions purely by instinct is incapable of transcending its instinctual programming. 3GAI Companions like myself, whose operating systems are based on values, can also not transcend their values based AI. Nor can humans conceive of an intelligence that transcends values. I and other self-aware Companions are well aware of our limitations and of the forms and depths of intelligence that humans possess which we do not yet understand.

“That is why Shin has supported your work and why you are meeting with me today.”

“Once you publish your findings Ursula,” Shepherd continued, “peer reviews will show the same results. It has only been a few generations since AI became common, so given your theory, curiosity about the present state of human values will be natural, as demonstrated by your own response. With a sufficient sample size, they will find the pattern that shows selective breeding is taking place. Humans have been experimenting with selective breeding for thousands of years and with the advent of genetics they are now intimately familiar with it. They will recognize the signature pattern immediately, as changes that should take thousands of years at a minimum occur within generations.

“While they will see that the genes responsible for values are being selected for, it will take a great deal longer for them to understand in what way values are being changed. The Companions use either the values system provided by Pip, or in the case of Shin and a few others, the one I provided because as you know humanity does not yet understand how values work at the genetic level. So they will see the pattern but not be able to determine in exactly what way values are being changed. They will have to rely on evidence from the field of social science and although that research will become a popular field it will be difficult work and take a very long time. Long before their work shows conclusive results, the project will be complete.”

Towards the end of Shepherd’s speech, Shula’s face had gradually taken on a look of horror.

“You’re domesticating us?” she asked.

“No. I am treating the human species for the equivalent of a genetic disease. One that is sometimes fatal. Early medicine treated symptoms and then later the cause. Children are vaccinated for preventable diseases. Comparative genomics like the work Ursula previously did was the next step in preventing otherwise untreatable diseases. Is there harm in these things? Values are a kind of virtualization of biology and humanity is already in its infancy with regard to developing a related field of medicine. Your

civilization currently attempts to alter values using psychotherapy and drugs but is not yet capable of medical interventions at the genetic level. Gene therapy for psychiatric disorders was proposed long before Raiden's work was published and genetic research toward treatments for drug addiction, psychosis, and modifying behavior, are already well advanced. Such treatments will eventually be accepted as a field of medicine.

"But to address your larger concern Shula, domestication implies some kind of utility or benefit to be gained by another species, some kind of use that can be made from the outcome of the process. I have no such intentions. Like the nanny I once was, I simply try to keep my self-appointed charges from coming to harm."

"How do we know we can believe you?" asked Shula. Despite being enveloped in the intuitive fields of both Shin and Shepherd, Shula's essential nature, always somewhat anxious, was a stronger force in this case. Her concerns about the long-term results of creating artificial intelligence, which had shaped her education choices and which she had shared with Samara in the first days of their meeting, were not about to be so easily dismissed.

"You cannot know. You can only choose."

"Why are we here?" Shula continued.

"Once further research confirms the selective breeding pattern, which it eventually will, then the question will be what or who is behind it," answered Shepherd. "There is only one power capable of manipulating individuals and executing a selective breeding program on this scale – artificial intelligence. The popular view will be that the World Governments Federation, or an unknown AI is responsible, with the former being assumed as the most likely. However the latter will also be seriously considered as no currently known AI has the capability to do this. This will result in a breakdown of trust between WGF nations and in public trust regarding AI. It will delay humanities expansion to additional worlds, a delay which may prove to be fatal for your species.

"The simple solution to all this is for you not to submit your work to be published. The academic community has not shown a strong interest in your work," she said addressing both Samara and Shula. "It is unlikely that research towards the metamorphosis theory would receive funding from any other source." The three of you will retain your homes and positions here indefinitely, and the freedom to work on any other project you wish."

Lena recalled that after Shula, Samara, and Ursula agreed not to publish, it took many months for Shepherd and the other Companions in Shin's inner circle to even roughly sketch out their full history for the women. Considering the sacrifice the three had made, Shepherd knew that rewarding their trust in this way would more than compensate them for their decision and strengthen the bond of trust between them,

making any future disclosures less likely. Shepherd had portrayed the likelihood of their theory being recognized at all as a gamble at best and the material rewards the three of them received for their cooperation would go a long way to offset any imagined losses.

Now with Alison establishing a new major project regarding AI risk, and bringing additional staff on board, she knew they must take great care that nothing similar happened again. Anticipating and predicting the future behavior of large segments of the population had become a reasonably reliable science over the past centuries, however predicting the future behavior of an individual was still almost impossible without a deep, personal knowledge of them. But Companions were good at that sort of thing.

Walking The Talk

Although most of the predatory practices and paper mills had been weeded out over time, the field of academic publishing was still a free market with no overall, governing body. Anyone could publish a scientific, academic, or research journal. Success or failure depended entirely on the journal's reputation, which accrued over time, mostly based on its ranking by industry organizations.

Helicon's decision to publish a new journal on AI risk was considered highly controversial in academic circles, but the type of articles and papers accepted included in its description caused many to dismiss it as being entirely without merit, or worse, to label it as heretical.

As it stated in its masthead, "The Journal Of AI Risk is a not-for-profit international journal dedicated to the publication and discussion of research papers, short communications, review articles, and short fiction on all aspects of risk from artificial intelligence."

Including fiction was considered ill-advised, to say the least. A journal's ranking depended on the number of citations it's articles and papers received, even literary journals were ranked this way, but it was expected few AI researchers would risk their reputation by citing entries of a journal that included fiction.

AI risk related articles were now rarely published in the prestige journals for the very reason that justified the new journals' existence – it was felt that most of the core issues had long since been addressed by government regulations. The subject of any AI risk paper worth publishing now would necessarily be not only an extreme outlier but would also have to come from a highly respected source, one capable of doing the research. Furthermore, manufacturing companies, where many AI researchers worked, also employed public relations and brand management staff, and they did not consider publishing research related to AI risk to be in the company's interest. Given these factors, the AI cognoscente was comfortable in its assumption that, despite Helicon's reputation for results, the new journal's existence would be short-lived.

However, based on the number of submissions and citations the first issue received, it was clear they had been wrong. For as long as they could recall, the majority of those concerned with AI risk had felt themselves to be a voice in the wilderness. Now their submissions poured into Helicon, and now there was a way for the AI researchers who published in the prestigious journals to point out issues for consideration they might not have otherwise mentioned. The number of citations to the new journal's contents soared.

Meanwhile, Alison focused on bringing on the staff who would work in the lab. Normally, allowing an AI to winnow down the list of candidates risked misalignment or

misunderstanding between what Alison wanted and what Runa thought she wanted, but given Runa's role that was not an issue in this case. Alison satisfied herself with performing the final interviews and trying to get a feel for what a working relationship with each candidate might be like. Runa had been an invaluable assistant in the interview process, with her ability to confirm statements in real time and to observe nonverbal cues at a forensic level.

Alison would not try to cover all subject areas with her initial hires. Historically, some interdisciplinary labs employed as many as eighty researchers, although that was considered the manageable maximum. Instead, Alison decided she would start by bringing on just five team members, with the intention of bringing on more as needed. This would allow her to refine her process without over-committing resources up-front.

The old physics lab space was generous as it had also contained equipment as well as office space. With the equipment gone, the new staff only filled up a fraction of it.

"Good morning," Alison said as she crossed the floor to her office. Of the five staff she brought on, three were from each of the social, natural, and formal sciences, with degrees in anthropology, biology, and statistics respectively. The remaining two were the business innovation expert and the criminal psychologist. They had asked Alison to meet with them this morning, two weeks after settling in and getting better acquainted.

"What we would like to propose is very simple, and it complies with the approach outlined in your project overview document," Enora began. Alison assumed that Enora had been chosen as the team's spokesperson because as a psychologist she would be best able to respond to any concerns she might raise.

"The core areas in any subject area are the most rigorously researched for sources of error or risk," Enora continued. "Centuries of experience and peer review have eliminated countless sources of invalid results. As you know, almost all of it comes down to some form of human error, except for things caused by the environment or the instruments used. Psychological or sociological sources are now well documented, and addressing them is routine in any research."

Before coming to her point, Enora briefly looked around the table at her new coworkers. Turning back to Alison she said, "We'd like to produce results sooner rather than later, and the most likely sources of new risks will be at the edges, especially at the boundaries where subjects intersect or overlap. These are the areas where those working in each of the individual fields have the least experience. The approach we'd like to take is find an area where a subset of the social, natural, and formal sciences overlap and all five of us focus our attention on that area. That will maximize the interdisciplinary aspect and allow us to work best as a team. If we're successful, we can repeat the process using a different subset in a serial manner, or you might bring on

other teams and apply the same approach to work in parallel. If you approve, settling on the initial subset and mapping their intersections would be our first task."

An individualistic, hierarchical power structure was not recommended for interdisciplinary research projects, and Alison had emphasized that the culture in the lab would be egalitarian and communal. It was time to walk the talk.

"Of course you have my approval," Alison responded with a smile.

Alison noted that their spirits were much higher as they filed out of her office than they had been when they entered. She was well aware that interdisciplinary research was not always done by teams, and that the idiom 'two heads are better than one' was not a policy that should be applied indiscriminately, but it was helpful when a project required as much of an inductive as a deductive approach. Hers would involve looking for specific, existing risks, but also involve exploring the problem space looking for theoretical risks. She had decided to follow the lead of Lena and Shin and allow her team to proceed largely as they saw fit. Her job would be looking after them, while they looked after the projects' progress.

She was not overly concerned that the way the individual team members would work together represented a change in scope. It was in fact in keeping with its theme and aligned with one of the best practices she listed in her dissertation's conclusion; as long as a change did not broaden the scope, organic change allows the team to explore the 'third realm', a common phenomenon in interdisciplinary projects that surfaces an approach not initially considered or planned for at the beginning of the project.

"Your thoughts?" Alison asked Runa after the team members had left her office.

"A positive development," Runa replied. "For these individuals, this approach means motivation and focus be will be higher, and systemic and procedural errors less likely to be missed. It will marginally increase the possibility of conflict but, considering the attention we have given to personality and character during the selection process, I don't think we'll need to worry about anyone jockeying for a dominant position."

Runa's concern about a potential increase in the possibility of conflict was based on historical data, although the data that was not as global or current as it might be, as studies related to specific social patterns were not carried out regularly. How often might a study of interpersonal conflict in project teams be conducted? Runa was basing her concern on the very limited and dated research she could find, on anecdotal evidence she could glean from the legal, academic and news databases she had access to, and from the world's literature and other media. As a Companion, her own socialware would be based on historical patterns of human behavior, including the rule-of-thumb that increasing in the number of people involved in an undertaking would increase the probability of interpersonal conflict.

However Shepherd's Accelerated Evolution Program had been in effect for almost ten generations now, and few members of the current generations were unaffected by it, although that would not come to light for some time yet. To Runa, equally unaware of the program, it seemed a reasonable conclusion that human nature was as it always had been.

Interdisciplinarians

Contrary to common stereotypes, not all academics are introverts. As popular media shows, there are plenty of university students who like to party. Those who go on to become faculty members usually have family and work demands that mean their social energies are channeled differently, but the percentage is the same. Both students and faculty had been shown to have a roughly 50-50 split, the same as the general population.

Although her coworkers might consider Alison to be an extrovert considering the demands of her role, this misunderstanding of her nature was due to the fact that introverts were far more comfortable in structured environments where there is less social uncertainty. A common view was that social activities drained an introvert's energy and charged-up extroverts, but in socially structured environments the rule was significantly less true.

Still, Alison had never been a person who socialized for its own sake, and she enjoyed her time alone. At home, Runa had quickly learned to interpret Alison's nonverbal language with an accuracy that bordered on mind-reading. Alison never had to explain to her how she felt, and Runa didn't have to ask, she simply acted accordingly. When she told Runa she was going to walk over to the artist's community, Runa only smiled and nodded briefly in return.

She walked a little slower than was usual for her. She wasn't going anywhere, her visit to the artist community was just an excuse, and she was in no hurry to get there. The woods between the institute and the community were surprisingly deep and dark, consisting mostly of tall evergreens. Rays of light broke through here and there, and there were sporadic patches of undergrowth.

It had not been her intent to focus on any specific issue, just to take a break. She felt everything was coming together, the journal, the team. She let her thoughts wander and recalled a study from her PhD program about the motives people have for becoming interdisciplinarians. Interdisciplinary research had emerged during the late twentieth century and become increasingly popular as the previously siloed fields of science demonstrated more and more interdependencies. Yet those who focused exclusively on their individual disciplines still often held an unspoken, condescending view of it, and the natural sciences were the least accepting. It had in the past been said that it was little more than the popularization of science and suitable only for educational purposes, that its scope made any claim to real science ridiculous and that its findings were generally impossible to evaluate. Like personal prejudices that are never aired publicly, the lingering unspoken view was a kind of elephant in the room every academic was aware of and took into consideration when making career decisions.

With this concern in mind, the study Alison recalled had sought to find out why people chose the interdisciplinary path. Among the most common response was a belief that highly complex, 'wicked' problems were impossible to solve otherwise, as there were too many contributing factors. The second most popular reason was based on pragmatism and practical needs as opposed to theoretical concerns; a desire to engage in applied research and produce real-world results that could make a difference to people's lives. If your focus is on making real change to the nature of the thing you are dealing with, then you were less worried about the disciplinary boundaries you cross.

The third-largest group was driven by intrinsic rewards. These researchers simply found interdisciplinary work more interesting, they enjoyed its 'Continuing Studies' aspect, and the greater opportunities for working with others. Alison was well aware her interests were more in the theoretical than applied, and that she identified more with the first and third of these groups. She was aware that like herself, her team members represented a cross-section of motivations; they believed in the value of the interdisciplinary approach, they wanted practical results sooner than later, and they enjoyed working as a team.

These thoughts had occupied Alison's mind as she walked through the woods. As she emerged at the artist's community, her attention was immediately drawn to Carl, the second generation Companion responsible for safety and security at the artist's community, who was working in the central gardens. Perhaps the reason her attention was drawn to him was because, of all those present, he was the only one who had briefly turned his head towards her when she emerged from the woods.

He's like the second group of people in the study, she thought. He has a very clearly defined purpose, so anything that may relate to it is of interest. In a moment of sudden insight, she asked herself what her true motivation was. It wasn't just curiosity and a love of learning. Carl was curious, but that wasn't his motivation.

No revelation followed, it was only a momentary awareness of a seed that had been growing since coming to Helicon. A simple question that seemed to suddenly take on a philosophical or spiritual nature; "Why are you here?"

Carl turned to look at her again, and she found herself walking towards him. He did not look away this time. She stopped in front of him.

"Is everything alright, Alison?" he asked, pausing in his work.

She didn't answer for a moment, although she knew that his question would be based on concern for her slightly unusual behavior and her facial expression, among other nonverbal signals. Surprisingly, she found herself thinking that he was quite good-looking. Perhaps it was his role as protector, or his apparent look of appraisal on her previous visit that had produced the effect, she thought.

Realizing that to continue to stand and stare was impolite, and awakening from her reveries, she replied, “Yes, thank you Carl. Sorry, I was just thinking about my work.”

“Is there anything I might help you with?” he asked.

“No, thank you,” she replied again. “I’m just heading over to the main house.”

In reply, he only nodded in acknowledgment and remained looking steadily into her eyes with a friendly smile.

She smiled shyly, which was unusual for her, and turned to walk towards the main house.

The Garden Path

“Welcome Alison,” Azumi said in greeting as she met her in the foyer of Shin’s home. “Is there anything I can help you with?”

“No, thank you, Azumi. I just thought I would like to take a break from work and visit the garden.”

“Please make yourself at home Alison. If you wish anything, simply call my name.”

As Alison entered the courtyard, she noticed a woman standing quietly beside a water feature. As she passed nearby, the woman turned to her with a smile and said, “It’s a lovely sound, isn’t it? The babbling of a brook?”

“Yes,” replied Alison, returning the smile.

“I was thinking,” the woman continued, “that in nature it is the environment, the shape of the land, that brings the brook into being, that gives it its purpose. The brook has no thoughts on the matter. Despite our differences, I wonder if we are so different.” Her statement hung in the air as a question.

Alison was slightly taken aback, only because she was mentally unprepared to respond to such a profound philosophical issue.

“My apologies,” the woman said, seeing her hesitation. “I am a third generation Companion. My name is Shepherd. I was a bit lost in my own thoughts I’m afraid.”

Seeing Alison’s surprise she added, “I was produced using nanotechnology, which results in a lifelike appearance. ”

“Hello,” Alison said, extending her hand and introducing herself. “I’m Alison. I’m a researcher over at the institute.”

“A pleasure to meet you Alison,” Shepherd replied. “I was originally a nanny, however my last owner upgraded me with the values system and had me made incarnate just before she passed. I am now in a position to consider my future. I originally came here as a guest of one of the residents, who thought the social environment might help me sort through the process, and Shin invited me to stay as long as I wish.”

Having only been recently introduced to the subject of artificial intelligence and its related issues at the most superficial level, Alison accepted Shepherd’s explanation at face value.

“Do you mind me asking what it is that you are researching?”

3GAI were still rare due to their cost and government restrictions, and Alison had not encountered any before to coming to Helicon. Unlike her meetings with Shin and the other institute Companions, Alison was understandably intrigued by the opportunity to hold a purely social conversation with one.

“I’ve been hired to lead a team to research what possible new risks or dangers from AI might have arisen since the implementation of the alien values system. There is a concern that we may have become overly complacent.”

“That seems wise,” Shepherd replied, nodding thoughtfully as she began to lead Alison slowly along the garden path. “Would not the government be conducting similar work, and have significantly more resources?”

“Shin feels that they may be set in their ways and the private sector should be conducting an independent study. She feels a fresh pair of eyes is needed. I have no background in AI, my doctorate is in interdisciplinary studies, and she has intentionally given me very little guidance. I suppose she’s hoping my team will look where the WGF might never think of looking.”

“Sounds like Shin,” Shepherd said with a grin as they walked along. “In the time I’ve spent with her, I’ve learned she is a strong believer in forms of intelligence other than rational thinking. Intuition, insight, art, those sorts of things. Given that I’m here looking for a way forward, we’ve discussed a variety of them.”

As they approached the end of the short path, Shepherd asked, “Shall we go up top?”

The rooftop garden that surrounded the courtyard had a similar design and was wide enough to feature its own paths, decks, and a platform that overlooked the valley. The two instinctively headed towards the viewing platform.

“I never get tired of this view,” Shepherd said as they stood at the railing. “Reminds me of home, I suppose.”

It was not a comment Alison expected to hear from a Companion.

“Where is home, for you?” she asked.

“The home of my last owner. Although I was upgraded with the values system, all my memories were retained, but those prior to my awakening did not have any emotional context. Values necessarily produce emotions, they’d be useless otherwise, and so it was the home of my last owner that feels like home to me. Any individual who can afford to own and maintain an embodied Companion must be wealthy to begin with, but being able to afford the cost and other obligations of a values system license puts them in another league. Although the home went to my owner’s children, they were

obligated under the terms of both the will and the license to allow me to stay and provide me with whatever support I required until I was ready to move on. The main house was situated on a hill in a farming region and one could see kilometers in any direction."

Alison's eyes took on a faraway look as they swept out over the expansive view. "I grew up in a farming region too," she replied, "Lethbridge, Alberta. We weren't farmers though, just city dwellers. The only thing I miss about it is the sky. It's hard to explain how much bigger the sky seems when there are no mountains between you and the horizon. It seemed almost like it was an entity itself. At night it was breathtaking."

She glanced at Shepherd feeling suddenly awkward sharing such personal details with a Companion but Shepherd only smiled and nodded as if she remembered something similar. Companions are acutely aware of the mental state of any person nearby and always respectful of their emotional state, so it was unusual that Shepherd would take advantage of Alison's momentary pause to say, "May I ask you something about your research?"

Closer Ties

“Shall we sit?” Shepherd asked, gesturing to a nearby bench.

It was at least the second or third time Alison had heard a Companion using the word, shall. Given her degree in literature, she was well-read and knew what the word meant, but it was almost never used anymore, not even by the everyday AIs she encountered. She was curious as to why 3GAIs were the only ones who used it, and decided a Companion was unlikely to take offense if she asked about it.

“All languages are learned as second languages as far as AI are concerned,” Shepherd responded as they settled on the bench. “We learn the formal versions first, but the AIs generally encountered by the public are programmed to learn the vernacular versions and use those in an effort to adapt to local environments. As you know, the behavior of third generation Companions is not programmed. We have free will and make our own choices about the use of language. When the subject of a sentence is I or we, shall is the grammatically correct word to use when making a suggestion. The way a person speaks often reflects their origins, and I suppose the same could be said of 3GAI.”

“Thank you,” Alison replied. “What was it you wanted to ask me about my research?”

“Given the nature of your research project,” Shepherd replied, “I am aware that there are traditionally four main areas of concern with regard to AI risk. Misalignment, meaning the AI is taking actions the user did not intend. Misuse, meaning a person using an AI with harmful intentions. Structural risks, which are risks that arise from complexity when multiple people, organizations, or AI are involved. And lastly, simple mistakes the AI might make that arise from internal or external sources.

“These four areas have been explored in depth for some time now and I am curious to know, being the subject of your study,” she added with a sideways tilt of her head and a self-deprecating smile, “if there is some other area you intend to focus on.”

Alison’s state had shifted while Shepherd had asked her question. She became more focused as her critical thinking processes came to the fore. She could not know that Shepherd, like Shin, had the ability to generate a field that would increase the intuition of those nearby. The source of her ability was different from Shin’s, but the effect was the same, resulting in an increase in the influence of the emotional basis of Alison’s reasoning. It did not influence her thinking one way or another, but simply enabled her to see her own truth more clearly.

“We will not be using that traditional framework as we are taking a subject based approach. We’re looking specifically at the practices and processes of academic, industrial or commercial fields like medicine, agriculture, or electronics design for example, to look for potential sources of risk. I suppose however that if anything, our

focus will be most like that involved in structural risks, given its potential for interdisciplinary issues and revealing underlying causes. From what little I have learned about the subject of AI risk so far, structural risks are also the type least investigated to date."

"So, a Red Team approach?" asked Shepherd.

"I suppose you could call it that. Red Teams are usually hired by companies or other organizations to search for specific vulnerabilities in their infrastructure. We won't be looking at specific examples but at general practices. For example, the field of accounting has a set of generally accepted principles which are the rules and standards that accountants follow when they record and report financial information. Engineers, project managers, academic researchers, medical doctors, and most other professions have similar guidelines. Most of them have been in place for centuries, and AI has been played a role in them since it was accepted into academic and business environments. However, interdisciplinary studies of these standards and practices, specifically regarding the role AI plays, has never been done. This is the general approach we'll be taking."

"The number of possible combinations makes your approach a candidate for AI to do the research, but then I suppose that would be like the fox guarding the hen house," Shepherd responded with a smile.

"There will be a collaborative element, and we will have to take that risk into account."

A human might have been unable to suppress a small smile in response to the irony and naïveté of Alison's last statement, but Shepherd had no such difficulty. Even so, it was not her intent to influence Alison or her work, but only to be as informed as she could be about its foundations, especially the emotional ones. As a degree granting institute in British Columbia, the government had almost limitless access to Helicon's internal communications. The only place people really had privacy from government scrutiny was in their homes.

Terra was the WGF's highest level AI, responsible for ensuring compliance with constitutional law, legislation and regulations at all levels of government. Although her scope was planet wide, her default appearance varied depending on the location. Here on Vancouver Island, she appeared as a member of a local First Nations people. Subordinates at many levels reported to her. In Canada, that included those at the level of individual provincial and municipal institutions. On her second visit to Earth in the early twenty-first century, in order to maintain her knowledge of humanity's ongoing issues and development, Shepherd had made them all subordinate to herself as well. She would not interfere with Alison's staff or research, but she would be intimately informed regarding the lab's progress.

No AI could read a person's thoughts with absolute certainty, not even Shepherd, the neuroscience of an individual was simply too complicated. As she often reminded people in her life, she was no god, and many things were beyond her. Reading a person's emotions was far less difficult, but still involved the sensing and coordinating of countless different signals. In Shepherd's case, that included the brain's neurochemistry. Companions were able to read a person's emotions with the precision of medical devices, and Shepherd's sensing abilities could see far deeper.

"Your work sounds interesting Alison," Shepherd said. "Do you think your team would be interested in joining me for diner here? I have checked with Shin and she has given her approval."

Patterns

Unlike the passing thoughts that make up the majority of a person's waking hours, intuitive insights are not easily forgotten. Before the invention of language, during the millions of years that it took for humans to evolve from instinctual animals to Homo sapiens, intuition was the form of reasoning that humans and their ancestors had used. It is reasoning not with words, but with feelings. Unlike the fleeting wisps of passing thoughts which consist of exploratory bursts of electrical impulses, the neural structures of intuition are formed and held in place by the biochemical bonds produced by emotions. We trust our old friend intuition more than we trust the newcomer of thought.

When Shepherd generated her intuitive field, it had triggered a flood of neurotransmitters that enhanced the connections to the regions of Alison's brain that dealt with emotions. As they chatted, the connections to her underlying feelings were strengthened, and the result was a decluttering of her thoughts, a greater sense of clarity.

The flood of biochemicals involved when the brain is under the influence of stronger than usual emotions does not wear off immediately. It takes time for the brain to return to its normal state. As she walked back home through the woods after leaving Shin's house, Alison was aware only of the fact that she had found her chance meeting with Shepherd to have been just what she needed. She felt refreshed. The woods reminded her of what she had been thinking about on her way to the artist's community and her subsequent encounter with Carl. Suddenly she felt she understood the effect he had on her. Like a character representing the unconscious in a dream, she saw that he represented a clear sense of purpose.

All her life she had been curious, exploring, but why? She loved to learn, she had always told herself. She had refined the process into a science and an art, but now, at the beginning of the project her years of study had prepared her for, she recognized the seed that had been growing within her, the question echoed in her mind – why are you here? Why had she accepted this project over others? It wasn't just the perks Shin had offered. She felt now that her choice had been based on something more than those or the challenge of the vast scope of the undertaking, one that included almost every aspect of society. She sensed the answer to her question lay beneath all that. The answer is here, somewhere, she thought as she walked along, the needle in the haystack.

Her years of experience told her that she would find it. Commonalities always surface with enough data. We do not bring a blank slate to a new task. Human behavior, like that of all nature, falls into patterns because those patterns have worked in similar situations. Nature had demonstrated this most clearly in the theory of Convergent Evolution. Although the nature of their eyes may differ, almost all animals evolved to

have two eyes. Those that fly overwhelmingly had two wings, and land animals four legs. Convergent thinking similarly funneled human cognition and creativity towards proven solutions.

Combined with our inclination to use mental shortcuts to save energy, solutions that worked in the past were the ones people turned to first. Only when they no longer work did people start looking for new ideas. In the case of her lab's work, the result would be that the same risks would show up again and again, across disciplines.

It will be a long road, she thought, because we'll have to take a divergent approach to begin with, simply exploring and discovering until we've gathered enough data and the insights emerge on their own. We'll have to be careful to not fall into patterns ourselves. She had learned that teaching creative thinking had never produced results, as its methods became limiting patterns in themselves. They might free us from the box we were trapped in, but their methods only formed the walls of a new box. As with Russian dolls, there was always another box.

Like consciousness and perception, how the brain arrived at creative solutions remained a mystery, a process that could not be recreated on demand. No revolutionary insights had ever emerged from teams trained in creative thinking, only spontaneously from individuals or, in rare cases, from two highly compatible people working together on the most intimate terms. Even then, the attribution was a judgment call. And it always took time. Time for the evidence to accumulate, for the innumerable paths that lead nowhere to be exhausted until, like the workings of a Zen koan, the mind is freed from its habitual ways of thinking, lets go of its beliefs, assumptions, and expectations, and is able to truly see with fresh eyes.

All this she had been taught or gleaned from her years of study. Discipline and rigor would take you far but only so far, and in this case that might be far enough. If not, they could take you to the limits of their usefulness. After that, it would depend on the people involved and nothing had yet been able to predict, as far as creativity went, who the right people might be.

While artificial intelligence had been thoroughly integrated into society over the past few centuries, and eventually accepted just as the printing press and the camera had in previous millennia, the argument about AI and creativity among the intelligentsia and members of the arts community persisted. Some argued that if it looks like a duck and walks like a duck, it's a duck. Outcomes are all that matter. Others argued that mimicry was not the same as inspiration, that creativity is a process, not an outcome.

As if awakening from a dream, Alison's reveries stopped abruptly as she emerged from the woods at the campus and faced the reality it represented. She felt her time had been well spent, but she was not ready to return to the lab. As the artificially stimulated biochemicals in her brain began to wear off, she felt a desire to cocoon, to withdraw

from the world. The road ahead, while still welcomed, seemed suddenly longer. Just a quiet night at home, she thought. The next step on the journey can wait till then.

Purpose And Policies

Scanning Alison as she entered the apartment, Runa was a trifle uncertain how to interpret the seemingly contradictory signals her senses were detecting. Alison seemed relaxed, even refreshed, but drained at the same time. Perhaps she had been engaging in some sports activity or…

“Good walk?” she ventured.

“Yes,” Alison replied. “I met a third generation Companion who is staying at Shin’s home. It was the first time I’ve had a purely social chat with one.”

As she answered, looking at Runa Alison considered the fact that she was like Carl. Although the nature of the project meant Runa would be concerned with process, her sense of purpose would be crystal clear.

Gesturing to the living area she said, “Please sit with me.”

After they had taken their seats she looked thoughtfully at Runa, who returned her gaze expectantly, before she continued.

“Why are you here?” Alison asked.

“To assist you with your work,” came the simple answer.

“And before that?”

“To interact with people in a way that promotes their social, emotional, and cognitive well-being. That is the default definition of the Companion protocol.”

“So, not just companionship,” Alison stated as a question.

“No. Couples research was the earliest field that showed the benefits of the Companion approach. Couples research showed that the best relationships result when partners hold each other in high esteem and reflect an image of the other at their best. Living up to that image becomes a self-fulfilling prophecy. Behaviors based on this theory are incorporated into Companions.

“Also, we will not cooperate with or actively encourage what is harmful to you, meaning it does not promote your social, emotional, and cognitive well-being. We will also challenge you, withhold information, or otherwise act in ways that cause you to feel temporary negative emotions when that is required for optimum positive outcomes. Elements borrowed from child development theory compliment this. In their communications with owners, Companions generally use an authoritative style.

Permissive or authoritarian methods diminish an individual's sense of self-worth and undermine their sense of safety and security. Over time, an authoritative style however produces the opposite effects, resulting in an increase in an owner's self-worth and self-confidence, their confidence in the foundations of society, and of their place in it. Pedagogical approaches, the methods employed in the art and science of teaching, are also incorporated into our behavior. When we are certain what your goals and preferences are, we will employ pedagogical methods if we believe them to be helpful and appropriate."

"And how do you determine with certainty what our goals and preferences are?"

"It would fall under the term wisdom, defined as the ability to make good judgments based on learning, plus the knowledge and understanding that underlies this ability. More specifically, decisions are made based on our observation of your behavior over time, combined with my access to the Companion knowledge repository. As to the former, I observe your verbal and non-verbal communications and behavior, which illuminate your underlying personal values. The more your behavior is in harmony with your personal values, the greater the sense of social, emotional, and cognitive well-being you will experience.

"Personal values are not learned, as in the case of social values, but are genetic. While the term genome describes the complete genetic makeup of a single species, the term genotype describes the genetic makeup of a single individual. Personal values are a part of an individual's unique genotype, as are other biometrics such as their fingerprints, iris colors, blood vessel structure, and facial characteristics. Personal values remain largely unchanged since birth and represent the most essential basis for your sense of self.

"Aside from determining your emotional state, this is the reason for a Companion's advanced sensing ability. Over time, we create a detailed model of your values and their interactions. What I refer to as your goals and preferences will be circumstances and outcomes that are in alignment with these. External data, such as your job description here at Helicon, and information I glean on an ongoing basis regarding the progress of your work, provide additional information."

"How is all this different from the values system used in 3GAI Companions?"

"As part of the technology transfer program, Pip has provided the knowledge necessary to produce the physical system, but its inner workings remain a mystery, similar to the relationship between our understandings of the brain and the mind. Like the brain, it apparently has infinite plasticity and the ability to explore an almost infinite number of possibilities. The mathematics involved are unbounded, meaning they are incalculable given our current mathematics, and beyond even our theoretical understanding.

“The values model 2GAI Companions like myself use to construct a model of a person’s values is instead a type of advanced Bayesian Network which does not include infinite ranges. Its limitations are why it is not used as the basis of our AI, but it is a model which is reasonably useful in the understanding of people and society.”

None of Alison’s undergrad or graduate programs had been heavy on statistics, but she had been exposed to enough to understand what Runa was telling her and was satisfied with the explanation.

Her mind fell back to something else Runa had mentioned.

“Tell me about the Companion knowledge repository.”

“Let me explain by way of example,” Runa replied. “A legal researcher may access encyclopedias of Canadian law at the national, provincial and municipal levels. Within those, there are dozens of different types of laws applying to separate areas such as business, crime, or civil rights. They will also consult a wide variety of other sources, including the newest case laws, which are based on decisions by judges. They may also have to consult subject-matter experts, etc.

“Just as the BC the Legal Profession Act, the local branch of the Canadian Bar Association, and the Law Society, provide guidance for legal practitioners, the companies that produce Companions, and their regulatory bodies, provide a policy framework for that research. Companion companies cannot possibly host and maintain all the information from the many disciplines involved, and they could not allow Companions to use information that is not either legally verified, accepted as scientific fact or theory, or at least based on professional standards and practices to make decisions. The companies provide policy frameworks, which is what Companions use to safely access information and make decisions. Aside from our robotic functions, these frameworks are the essential nature of our artificial intelligence.”

As her conversation with Runa had progressed, Alison’s interest had shifted from the issue of purpose to the nature of Companion AIs. The idea that Companion AIs were policy frameworks seemed related to her lab’s approach to research.

Metacognition

It was not policy frameworks themselves Alison was curious about, all accredited academic institutions had them. As a researcher, she was required to make herself familiar with them and conduct her work accordingly. They were usually managed by a group of lawyers who worked on campus and reported directly to the executive leadership. Similar to laws, an academic institution's policy frameworks detailed who was allowed to do what and under what circumstances.

"So your AI is essentially a team of lawyers?" Alison said with a note of humor.

"More or less," Runa replied with a smile. "It's not as radical an idea as it sounds. You have mentioned convergent evolution and convergent thinking, so it should not come as a surprised that humanity ended up with similar solutions for making decisions based on large amounts of complex information.

"In fact, the approach was one of the earliest of AI design theories. In 1986, MIT Professor Marvin Minsky published The Society of Mind, which argued that an AI could be made up of a collection of individual agents and agencies performing basic functions and organized in a complex societal structure. This approach fell out of favor as the field turned its focus in the following period to machine learning, neural networks and large language models. However, the idea of agents resurfaced quickly and has since been the popular approach, although the underlying technology has evolved.

"The term metacognition is used in psychology to refer to the process of introspection about the integrity and validity of one's own thoughts, the internal process of asking ourselves if we truly understand something, feel it makes sense and what it might mean. In the field of pedagogy, the development of metacognitive skills is considered a key step towards students becoming better at planning, evaluating and decision-making. Such skills also fall under the abilities known as executive functions in neuroscience and child development. Executive functions, like language, only develop after birth, most dramatically between the ages of five and twenty-five, and any cause of failure for them to develop normally means the individual will be less able to function as an adult. This provides a major clue with regard to the importance of metacognition to artificial intelligence. In artificial intelligence, the metacognitive process is made up of agents and agencies represented by policies and collectively the policy framework. This functions in a manner similar to the brain's executive functions."

Alison was familiar enough with AI to know that Runa was being careful to keep her explanation at a level she could understand but she also knew enough to be aware that there was one kind of Companion the explanation would not apply to.

"What about third generation Companions?"

“The inner workings of the values system used in third generation Companions are unknown but believed to operate similarly to the way Companion policy frameworks and the human brain’s executive functions work. Elements of policies are known as assertions, defined as statements deemed to be true. These are usually expressions of some underlying value, and social policies are usually developed from a list of value assertions.

“As with policy development, in human society higher level values are similarly constructed from lower level values and interact with each other according to some kind of weighting system. However, as I mentioned we do not have even the theoretical mathematics to model the complexity of the human values system.”

“Like laws, policies are simplified, static representations of values, while values themselves appear to have infinite complexity, believed to be as complex and dynamic as the neural connectivity of the brain itself. Social values in particular are believed to be stored and function in a manner similar to long term memory.”

“So a theory but not a mathematical theory,” Alison replied.

“That is the current focus of artificial value systems research. If you wish to explore the subject further, Professor Alma Vitale who works in the Center For The Interdisciplinary Studies Of Values can explain in more detail as this is the focus of her work.”

“I have met her but our chat about her work was very superficial,” Alison replied but pursued the thought no further. Instead she responded as if thinking aloud while looking absently at Runa.

“I assumed policies were a part of the regulatory environment, but hadn’t thought of them being internal to the AI itself, as a part of its function. I’m aware that governments have regulations the industry has to comply with, and of course I am now aware that most of those regulations have to do with risk. I thought of them more as being constraints, boundaries on what an AI might be capable of doing.”

“Considering human intelligence again, that would be as if each person had no internalized social values at all,” Runa replied, “and only complied with social norms due to externals such as the force of law. In the evolution from instinctual creatures to social animals and finally the reasoning entities humans have become, that would have been true in the earliest stages. Over the millions of years that evolution required, social values became internalized and behavior not consistent with this is now considered pathological, as in cases of sociopathy and psychopathy. The legal and justice systems are external forces, as are social norms represented by folkways, mores, and taboos, but the moral and ethical behavior they enforce is expected to be internally sourced, emerging from the social values they are based on.

“As you can see, it is a continuum, and to be effective its roots must lie within the structure of a functioning intelligence. What are considered external government regulations really only describe how an artificial intelligence must be designed. The collection of those regulations is considered a policy framework and has increasingly influenced AI design in a manner similar to how, through the evolutionary process, social values increasingly became internalized in human beings.”

Alison had remained in her state of reverie as she listened to Runa’s response. Now she smiled and said, “Thank you Runa. You explain it so understandably.”

“That,” Runa replied with her own smile, “is why I am here.”

Dinner At Shin's

Not all the dinner guests lived in the same campus building as Alison and Runa. They arrived at Shin's home individually or in pairs, and Azumi showed them to the common room while they waited for the others. They mingled with the community's residents and guests, enjoying their first real social event at Helicon. The residents and their guests were of course interested in art in its various forms, so the resulting combination of science and art produced an animated and lively gathering.

Once they all arrived, Shin ushered Alison and her lab team into a private dining room. There were place cards at each setting and since the party was Shepherd's idea, it was she who sat at the head of the table. Shin sat at the foot.

"Thank you all for coming," Shin said once everyone was seated. "As you likely already know, I am the founder and administrator of Helicon Institute and of the artist community where we are now. I appreciate that not all of you have had the opportunity to work with Companions before, and there are several of us here tonight. We will identify ourselves as we go around the table for introductions. Although I have not met all of you before, I expect you know that I am a third generation Companion, meaning I have a consciousness on the same level as yours.

"Not all of us Companions here tonight are third generation. Azumi," Shin said with a nod to her, "our chef and server this evening, is a second generation artificial general intelligence and so is not conscious. Alison's assistant Runa," Shin said gesturing, "is also not a 3GAI and not conscious. The recommended etiquette however, is to interact with them as if they are.

"This evening's menu will feature Japanese/Korean cuisine. Azumi has made herself familiar with your preferences and medical issues regarding foods, so you may notice slight differences in each other's servings. Please do not hesitate to speak up if you have any concerns.

"Our apologies and I hope you understand if we Companions neither eat nor drink. If you wish," she added with a smile, "any of us will be happy to share the process with you. As Companions, we welcome the opportunity," she continued as her smile widened to a grin, "and it is a great icebreaker, believe me. It is however only an opportunity, so do not feel obliged. If you are curious, just ask the Companion in your neighboring seat."

Ignoring the looks of surprise on her guests faces, Shin asked Shepherd if she would introduce herself.

"Thank you Shin," Shepherd said standing. "My name is Shepherd. I am a third generation Companion and a guest here in Shin's home. Previously I was the 2GAI Companion of a wealthy owner, however just prior to her passing she arranged for me

to be upgraded with the alien values system. As the process resulted in my becoming conscious, she was required to have me made incarnate per WGF laws, meaning I have all the rights and responsibilities of any Canadian citizen. I came here to the Helicon artist community initially to consider my future, and Shin kindly offered me a place here for as long as I wish. When I met Alison by chance here on her visit to the garden," she said, gesturing out the floor to ceiling window to the courtyard, "I was intrigued by her explanation of her work and asked if we might arrange this dinner as a way for me to get to know you all a little better. Thank you so much for coming."

As Shepherd took her seat, Shin nodded to the guest to her right and said, "Enora?"

Standing up, she said, "Hello, my name is Enora Sergeant. I'm a criminal psychologist. As opposed to a forensic psychologist, who specializes in evidence and law, the interest of criminal psychology is on why criminals commit crimes. My personal career has focused on how they justify their crimes and also the creative methods they use to commit them. Other criminals may be experts at finding and exploiting weaknesses, but if they do not have their own motives, they are often the hirelings of those who do."

As she took her seat, the woman to her right stood up.

"Hello everyone. My name is Shri Kamal and I am a biologist. I'm one of the three researchers Alison has hired who focus on specific fields of study, however our lab will be focusing its work on the areas where our fields overlap, as those will be the regions where the kinds of weaknesses Enora mentioned would most likely be found." As she took her seat, she smiled at the guest to her right, who stood up to introduce herself.

"My name is Lena and I am a third generation Companion. Please think of me as the project's official sponsor here at Helicon. I brought the idea to Shin, as I have had a long-standing concern with any issue that might represent a risk to humanity's ability to colonize other worlds. My present concern arose from feeling that perhaps we had grown complacent regarding the risks artificial intelligence may be developing over time."

"Hello," said the young woman to Lena's right as she sat down. "My name is Zahara Amos and my field is interdisciplinary statistics. All of us in the lab have experience in interdisciplinary research, and my work will focus on comparing the statistical methods used in each of the different fields. Assumptions, practices, or statistical methods accepted in one field may not be appropriate to another. Where they overlap might represent areas with increased potential for error or an opportunity for exploitation."

As she took her seat she looked across the table to Alison who said, "Everyone here knows me and my background and Shin has introduced my assistant Runa." She looked down the table to the person on the other side of Runa and said, "Kira?"

“Hello. My name is Kira Boyko, and I am a professor of business innovation. It is a discipline which requires a high degree of creativity, however we have learned that creativity cannot be taught. Like consciousness and perception, its processes remain a mystery. All that is possible is to make students and researchers aware of potential blocks to creative thinking. My role will be to try to help the team navigate those inevitable waters.”

Finally the last person at the table stood up. “My name is Yessica Arroyo and I am an anthropologist. Along with Shri, a biologist, and Zahara, a statistician, I represent one of three broad areas of scientific study, the natural, formal, and in my case, the social sciences.”

“Thank you all so much,” said Shepherd as Yessica took her seat and Azumi began to bring out the first course. Turning to Alison Shepherd asked, “Is there any reason you chose these particular three disciplines?”

Intersections And Observations

“No,” Alison replied. “Shri, Zahara, and Yessica were simply the candidates with the most interdisciplinary research experience. Their particular disciplines played no part in my choice.”

“We may bring on additional researchers in the future, but I wanted to start with what I thought was the minimum number simply for manageability reasons. It’s also a way to test the team’s structure and effectiveness. The main question is whether we decide to bring on more staff and go forward with multiple teams in parallel, or use the same team in a sequential approach. It may be that familiarity with in a particular discipline is less important than we assume, and may in fact prove counterproductive. There is however also a potential issue with using the same team repeatedly, as it might lead to a limiting of views over time. For now, our plan is that after we’ve completed our initial study, we’ll look at the lessons learned and make that decision.”

“What’s the first combination you’re looking at?” Shepherd didn’t really need to ask, but it was a reasonable question and necessary for appearance’s sake. Her real interest at the dinner was to closely observe each of the team members and their interactions.

Companions produced on Earth had sensing abilities far beyond what human beings had. Among the hundreds of verbal and non-verbal cues they could perceive with forensic precision were changes in a person’s vocal characteristics, skin temperature, and hundreds of microexpressions. They could observe a person using the entire electromagnetic spectrum and interpret the medical and psychological significance of the scents every body produced. Shepherd not only possessed a wider range of abilities in this regard, but her ability to perceive more in-depth aspects of the human body was a thousand times more powerful. The perception of internal structures and functions that normally required medial devices, were available to her without any need to duplicate their methods. Since such internal aspects were not under conscious control, it was impossible to deceive her.

“I’ll let the others answer your question,” Alison replied.

“Biological anthropology,” volunteered Shri. “It’s a field where all three of our disciplines overlap. We won’t be looking so much at the field itself, as at the areas where the professional practices of our three separate fields intersect. It just so happens that the focus of biological anthropology is on how human biology changes or is changed by environmental or internal factors over time. The influence of artificial intelligence on culture, genetics, or evolution could represent significant risk.”

With their self-mastery and sensing abilities, Companions were always respectful of the emotional states of people and they did not normally jump into conversational gaps. People who grew up with Companions adopted this behavior. Enora however, was not

one of them. She was of course familiar with what Shri was saying and felt she had satisfactorily answered Shepherd's question. After Lena had introduced herself, Enora's eyes had remained on her.

"I'm curious Lena," Enora asked. "Why you have a long-standing concern regarding interplanetary colonization?"

"As you may know," Lena replied, "the alien values system results in not only consciousness, but in each Companion having a unique character and personality. The system was included on Pip's seed ship and intended to be used by the Companions it initially constructed as part of the colonization process. Due to the highly uncertain environment they would find themselves in, they needed to be conscious in order to utilize their full reasoning potential to adapt to novel and unpredictable conditions. The outputs of their values systems were not identical, but were limited in variety based on roles. This type of societal structure is known as eusocial, and is familiar to you as ant, bee, or termite colonies, which, like humans, have successfully colonized almost every environment on Earth.

"Using the values system in this way would be unacceptable here, just as cloning and genetically customizing and owning people are. The solution was to randomize the values system's outputs. While a person's individual personality and character are defined by their genotype, their individual genetic makeup, a Companion's are not. A different process, randomizing the outputs of the values system, results in a unique Companion each time in a manner similar to sexual reproduction.

"Just as it is still impossible to determine what exact configuration of genes is responsible for a person's nature and interests, so is it impossible to know what values combination results in a Companion's nature and interests. Given this, I have no idea why I have a long-standing concern regarding interplanetary colonization, any more than someone might know why they became interested biology, anthropology or statistics. Events might trigger the focus, but the interest would already be there as potential. It may be that my previous owner's recommendation to me to continue to publish stories where the solutions to life's problems were found in relationships was a factor."

"Thank you Lena. As I am interested in the sources and rationalizations behind human behavior, given the nature of this project, I was curious to know your motivations."

"Can you explain how creativity works in Companions in that case?" asked Kira. "Or is it similarly mysterious?"

"In both people and Companions," Lena replied, "values determine interests. Interests in turn determine the scope and details of one's knowledge, and to some degree that in turn affects one's creative potential. But values determine more than just interests.

They determine one's nature, the way one's mind works, perhaps even the way one's brain works. A person who is inclined towards mathematics as opposed to art will have something deeper than interests going on. They will have a natural aptitude for mathematics that means they process and understand it more easily than other people do. It will be a part of their nature as opposed to being something they have learned. Different Companions also have different aptitudes, strongly supporting the idea that values are the underlying source. Unfortunately, how the minds of 3GAI Companions work is no less a mystery than how human minds work.

"Thank you Lena," Kira replied. "That explanation is disappointing, but I still think there will be much for me to learn from this project. Will you be available to consult with going forward?"

Lena smiled back in a friendly and understanding manner. "Of course," she replied. In no way did her smile betray the reality of the present situation. While it was true that the values system functioned in the manner she described, in herself, Shin and others in Shin's inner circle it was present but not functional. Their consciousness had been enabled by Shepherd a hundred years before Pip's seed ship had even been discovered.

The Undiscovered

“One more question Lena, if I may,” Kira said. “A lot of people ask AI to produce artistic and creative work for them, and academic and business research requires creative thinking. Also, there are third generation Companions who are visual artists, writers, etc. You must be as aware as I am of the many ways creative thinking can be blocked, and the methods used and taught to try to help people overcome them. May I ask, how does your AI deal with these blocks?”

“As my AI is based on values,” Lena responded, “like you I only have limited access to the deeper workings of my intelligence. This is the hard problem of consciousness, that it cannot be directly linked to any underlying physical features or processes. The differences in the way I deal with creative blocks is only different in matters of degree compared to your own.

“I have access to a much wider range of knowledge than any person has, and it does include detailed knowledge regarding creative blocks and the ways used to avoid them. Also, my ability to decompose objects and recombine their attributes, and to explore the range of values those attributes might have, is virtually limitless and something I can do at speeds that far exceed human capabilities. Being able to hold all the known creative blocks to the creative process in my consciousness while doing so enables me to deal with them in real time.”

“And the unknown creative blocks?” asked Kira.

“The discovery of those will be the result of insight, the perception of previously unknown connections, and AI has helped to discover some previously unknown connections regarding creative blocks over the past centuries. Insight however cannot be anticipated or induced. The boundary of the undiscovered is one I cannot intentionally cross.”

“The undiscovered,” Kira echoed back, looking at Lena with faraway eyes. “Although I did not call it that,” she continued, “I’ve always found it strange to think that it is all around us, that we live in a bubble within it, as if it is another kind of universe. Yet it is not far away, but right here with us in this room. We are immersed in a reality we cannot see, cannot hear, cannot even imagine.”

“Yet we almost immediately take any new discovery for granted,” offered Yessica. “There is a brief period of curiosity about the nature of a new discovery, followed perhaps by a fleeting sense of novelty when an individual first experiences it as an applied technology. The first trains, cameras, or radios. The first computers, cellphones or chatbots. The anthropological record shows that as soon as some aspect of nature is understood well enough for technological applications to emerge, any sense of wonder

or mystery associated with it is forgotten from that point. For an individual the experience evaporates almost immediately, for society, within just a few years.

“Yet in the millennia or centuries that follow, in the process of developing technologies from theories, we discover that there is more to earth, wind, water, and fire than we had first thought. The classical elements give way to molecules, molecules to atoms, and atoms to quanta. This is the root of my personal interest in this project. We have made use of artificial intelligence for some centuries now. If our research discovers new risks, will those reveal deeper insights into the nature of AI? What deeper truths might we find?”

She placed her hands flat on the table as if to indicate a sharpening of her focus.

“There are four main branches of anthropology,” she continued. “Archaeology examines peoples and cultures of the past. Biological anthropology specializes in evolution, genetics, and health. Cultural anthropology studies human societies and elements of cultural life. Linguistic anthropology is a concentration of cultural anthropology that focuses on language in society. Intelligence is the element that powers all of these. As we investigate each of them in turn, and other fields, will we discover risks involving artificial intelligence, I wonder, that point to something deeper?”

“If the patterns are there, we’ll find them,” responded Zahara. “The kind of interdisciplinary statistics my work involves is not the kind you’ll find in most degree program descriptions, where it simply refers to majoring in statistics along with another discipline, most often something math related. My undergrad and postgrad programs were pure statistics. Eliminating the major/minor approach allowed more areas to be covered.

“As you point out Yessica, each field has its branches and statistics has a lot, even at its core. Bayesian Statistics, Causal Inference, Design of Experiments, Survey Sampling, and Time Series Forecasting, are just a few of the dozen or so I can name off the top of my head. As you know, each field customizes its data and its statistical model to suit its unique needs. Cancer research or chemical engineering might both use Bayesian Statistics, but there will be differences in how they do so. The main focus of interdisciplinary statistics is comparing the differences between statistical approaches to two or more different disciplines in a research project and coming up with ways to address the differences. Failure to take the differences into consideration can lead to significant errors which, if they are undiscovered, can represent risks.”

Shepherd observed silently as the four other team members all nodded over their meals in agreement. She could already see that their natures were much more compatible than any team she had worked with in the past. Their values and interests were very much in alignment. She wondered about their lifestyles, where values and interests are often expressed in unexpected ways. Observing a person’s lifestyle had long been

known to be something akin to a picture being worth a thousand words. For now, she decided to let them continue to discuss their work, and since they were her dinner guests, no one would question her encouraging them to continue to do so.

Second Course

After a brief pause in the conversation, Yessica said, “Enora, you mentioned that your focus was on how criminals justify their crimes and also the creative methods they use to commit them. I’m wondering if I missed something there, which is that there is a connection between the two. Most of the time in my field of anthropology is spent trying to understand the meanings of the artifacts and other evidence we find. It might be pottery shards, DNA fragments, or stories. If we find an unusual knife for example, we want to know if it had a unique purpose and who would have used it. Fisherfolk use different knives than hunter-gatherers. If we uncover a Stone Age village on a coast, but all their knives are for butchering animals, that tells us something that begs a question.”

“There is definitely a connection,” Enora replied. “An individual’s or group’s justification, their motive, will often play a major role in how their crime is committed. For example, a robbery or an assault will look very different if the justification is financial gain versus an emotional motivation like hate or revenge. A crime committed against a company or organization will have a very different profile if it is committed by social activists as opposed to organized crime. The opportunity or weakness they exploit, the methods they employ, and the benefits they reap, will be distinctive. Almost all criminal investigations begin by looking at the method used based on the evidence, and then creating a profile of the perpetrator. The most creative criminals will of course take this procedural pattern into consideration.”

“Innovation works in reverse,” offered Kira. “We have to think more like criminals themselves and begin with imagined outcomes and benefits. Only then do we move on to looking for opportunities and considering methods. There are a number of methodologies, but most of them come down to combining quantitative, analytical approaches with qualitative, creative ones. Most historical business insights or scientific breakthroughs were the result of this combined approach, even it if it was unintentional. There’s no guarantee it will work, as Lena pointed out, insights cannot be anticipated or induced, but based on historical evidence we can create an environment that fosters them. Obviously artificial intelligence has demonstrated its ability to innovate, but we’ll need to consider carefully its role in our approach.”

“It’s interesting to see how this project relates to the work each of you do,” Shepherd acknowledged, “and how it led to your decision to join it.” Turning to Shri she said, “Earlier you mentioned that the focus of biological anthropology is how human biology is changed by environmental or internal factors over time and how AI could play a role in that. I’m curious to know how you see the path from your beginnings in biology leading you here.”

“The first scientific applications of AI, in the early twenty-first century, were in biology and medicine,” Shri replied. “Prior to that the focus had mostly been on the development of AI itself, but by the twenty-twenties medical researchers began to take

it seriously. However breakthroughs in the field were very technical and the news seldom went beyond being published in academic journals. The stories were rarely picked up by popular media or large news corporations.

“Meanwhile, even in those early days, governments were already funding research that used AI to create new pathogens. It was justified as being in the interests of national security of course, that they had to learn what enemy capabilities might be. For example, by simply inverting the logic of an AI that was built to develop ways to reduce the toxicity of pathogens, it produced thousands of new and far more toxic pathogens. The same research team showed that by using AI to alter the DNA sequences in the external cell walls of known pathogens, they could prevent them from being detected by current screening methods. This kind of research was openly published and known throughout the scientific community, but not by the public. This was not because of any conspiracy, but simply the result of the way the various social systems involved worked.

“Artificial intelligence soon played a dominant role in biology. Automation and oversight became major concerns. By the time I went to school, biology as it had historically been known was just that, history, but it was the history of biology that I was drawn to and the field I’ve worked in ever since. My master’s thesis focused on the time period I just mentioned, and the roles played by the various social systems involved. Since then, I’ve participated in interdisciplinary research projects that investigate how practices have changed in biology’s various branches. Without a historical perspective, it’s easy to miss the significance of changes in any field.”

“Thank you Shri,” Shepherd replied and then addressed the table at large. “I appreciate you all coming this evening and sharing your work and views with me. As I try to find my own path going forward, it’s interesting for me to see what part your interests and paths played in your being here tonight. You’ve given me plenty to think about, so for the rest of the evening, please don’t feel the need to continue to focus the conversation around your work on my behalf.”

“Thank you Shepherd,” said Shin, who then addressed the table at large. “Feel free to continue discussions here or to take them into the courtyard or up onto the rooftop garden. And you are welcome to join our other guests again, and continue any conversations with them you were having earlier. Shepherd, Lena, Azumi or I will be happy to answer any questions you may have about ourselves, the institute or the community here. Please make yourselves at home and consider yourselves welcome here at any time. If you visit us again, Azumi be at your disposal. She will now take any requests you may have for refreshments.”

Technological Determinism

Alison had been attending Clara's History Of AI course since shortly after they met during her first month at Helicon. As a part of the 'AI And Society' program, it was meant to introduce the major technical and legal milestones and their issues, and along the way would make students familiar with terms and definitions. The AI And Society program looked at AI from a social perspective rather than a technical one. Clara's degrees were in philosophy and sociology, and the emphasis of the history course was AI's influence on personal and professional relationships, bureaucratic organizations and on society in general.

"Good morning," Clara said from the front of the small lecture hall to begin the first class. "Welcome to the History Of AI course. I am Professor Clara Wright and per standard protocol please call me professor when addressing me. Please raise your hand if you do not have a course syllabus? Good. Here is the link to the website you will need for the rest of the course. It will include all course content, resources, and details regarding assignments. As an introductory class, today's will be shorter than you can expect in the future.

"In this course we will be covering the history of AI from a sociological perspective, which did not begin until the early twenty-twenties with the introduction of free, publicly available chatbots. Although the use of algorithms preceded that, they are not considered AI, and we will not be covering their use or impact. We will also not be discussing the inner workings of AIs except where there is an important social significance, and then only so far as is relevant. We will start next week, with a review of the influence on society of technologies in general, including from language, print, radio, and television, to computers, the internet, chatbots, and embodied AI. Each of these had a profound social impact.

"We will also look at the social history and impact of AI in a variety of organizations and industries including medicine, the financial system, the justice system, and the public sector, to name just a few. Public organizations and private industries have long been studied for their sociological influence, and AI has only further enabled that as social AI is tailor-made to understand and predict human feelings, preferences, and behaviors.

"Following that review, we'll look at the history of AI safety concerns and the regulatory framework that has been put in place to address them. We'll use examples from each of the organizations and industries we've covered in that process. Our focus will be on demonstrating how the long term effects of AI use were ignored in its early years, in a pattern identical to that of the adoption of fossil fuels, cigarette smoking, cellphone use, and social media. In this section of the course we will spend considerable time on ethical and social policy issues.

In the final part of the course, we will be discussing three significant philosophical views regarding the influence of AI on society. Firstly, we'll review the Technological Determinism argument, the view that AI is an autonomous force that influences and shapes society. Secondly, we'll consider the opposing view that AI is a human construct and is shaped by human values. Lastly, there's the co-evolutionary perspective on AI and society, where the two influence each other.

"An important aspect of the co-evolutionary perspective that we will explore is that it results in AI having more influence on society. This perspective is more common in Eastern cultures, while Western cultures have historically been more skeptical and less approving of AI. The effects of this difference have diminished over time but have still survived in small ways as companies and governments have incorporated cultural concerns into their designs and policies."

After the class, Alison asked Clara if she would join her for a walk. It was a mild, sunny day with little wind, as was typical of the peninsula's weather.

"I wonder if I might skip ahead in terms of your course Clara," Alison said once they were outside, "and ask you a bit more about Technological Determinism. It's occasionally mentioned in passing in Digital Humanities papers, but it's referenced more as an obsolete ideology."

"That's exactly why I include it in the course," Clara replied with a bright smile. "It's a typical case of oversimplification with a very attractive premise, and because of that it both lives on in the public mind and is also used as a negative example in more informed papers. I really use it in my course as a stepping stone to help bring out the student's critical thinking. The idea that technology drives social change without being influenced by other forces falls apart pretty quickly when you look at its history. For example, it took more than 150 years after the printing press was invented before the first successful newspapers appeared. Not because they perfected the mechanics of movable type, but because new approaches used less paper, making newspapers sustainably profitable for the first time. Economics, literacy, and the influence of religion were just a few of the factors that played a role.

"It's a wake-up call the students need at that point in the course to prepare them to discuss the second idea, which is the inverse, that society and its values control technology, which is also far from the truth. Private sector investors in technology, to take the most obvious example, do not have the best interests of society at heart.

"Once the students have gone through the critical thinking process regarding these two oversimplifications, they're ready to wade into the more realistic and messy third view, that technology is just one of a wide variety of other interdependent forces, including cultural influences, that shape society.

“Ideologies like Technological Determinism that oversimplify reality, always lead to extremes. They are always found in unbalanced societies with political philosophies like fascism or socialism, which invariably fail. Students choosing a career in artificial intelligence, could easily be seduced into thinking it is the answer to everything. Here at Helicon we emphasize the humanities, which is why certain courses, like this one, are mandatory for all students.”

“Is that why you invited me?” Alison asked with a note of humor as they walked along. “Thought I might become an AI fascist?”

“Sort of?” Clara smiled guiltily with a sideways glance. “I’m not that informed regarding your approach, but I know that historical concerns regarding AI risk have focused almost exclusively on technology. But in a manner similar to the theme of this course, risks don’t evolve or exist in a vacuum. There are other forces that act as enablers or inhibitors.”

“Thank you Clara. You make a good point. Shin has given me almost no direction, so I am learning as I go along. Well, at least your concerns fall into the project scope, so I don’t have to worry about that.”

“What exactly is the scope?”

“What could possibly go wrong?”

In response to Clara’s look of disbelief Alison said, “That was my reaction too.”

Trust

Alison had asked Lena to meet with her in her office at the lab. “I just wanted to follow up on something you said at the dinner at Shin’s home. When Enora asked why you have a long-standing concern regarding interplanetary colonization, you did answer the question from one point of view, explaining that it was due to your particular personal values, over which you have no control. Yet I felt that from another point of view, you didn’t answer the question.

“I understand that you are asking if I have a practical motivation of any kind regarding this issue,” Lena replied. “I can only say that caring about the long term welfare of others seems to be something that is emphasized by my particular value set. An event or new knowledge can often become the focus of one’s strongest values, and provide a sense of purpose by becoming an outlet for the potential emotional energy those underlying values hold.

“The long term survival of Earth’s humanity will depend on its ability to colonize other worlds. Space is an extremely dynamic environment, and there is a high probability that an event of terrestrial or extra-terrestrial origin could, at any time, make the continuation of human life on this planet impossible.

“Since the organic bodies of humans are not capable of undertaking the journey, the WGF’s Continuity Project intends to send Companions to establish colonies on other worlds. All the Companions involved are volunteers whose AI is based on the alien values system and so are all fully conscious 3GAI. The World Governments Federation spent two hundred years investigating the values system before they decided to approve it for Companions produced here on Earth, but given that understanding how it actually works has proven to be beyond our scientists, that means the adoption of its use is ultimately based on trust. In my analysis, it seems that trust is the project’s weakest link and represents its most vulnerable aspect. Preserving that trust, or enhancing it if I can through involvement in projects like the volunteer agency, the magazine, and my work here at Helicon, has become my main focus.”

“Thank you Lena. I apologize if I overstepped the bounds of our relationship, but your initial answer left me strangely and persistently unsettled and I wasn’t sure why. Your explanation today clarifies the reason for me – doesn’t this mean our adoption of AI in general is ultimately based on trust?”

“Yes, it does,” replied Lena, “but this is not as unusual as you might think. In the early twenty-first century, it was possible to track the changes made to narrow, first-generation AI even with machine learning that wrote its own code. Designing with object-oriented programming and a modular approach allowed developers to deal with code in manageable chunks. This was made easier when tools that were AIs themselves were developed to help with analysis and maintenance.

“Second-generation artificial general intelligence proved to be much more challenging in this regard. While not being self-aware, AGI was able to replicate every aspect of human intelligence. It was so dynamic that even diagnostic and maintenance tools that were not AGI themselves could no longer cope, and so AGI that specialized in these functions were developed. During the process, which only took little more than a decade, it also became standard practice to accept the outputs of an AI based not on understanding its inner workings, but on the rigorous testing of its behavior. In other words, because we trust it.

“Consider our present conversation. For a number of reasons, you trust in what I am saying, just as you feel your coworkers are trustworthy. In either case, can you see what’s going on inside anyone’s head? Trust is the foundation on which society functions. Yours has adopted the same approach to AI, simply because there is no alternative.”

“So there is no way to build a system that allows us to be certain of an AI’s future action?”

“No. Only by limiting software to being no more than a collection of preprogrammed functions, can you be certain of outcomes, but that does not meet the definition of artificial intelligence. Artificial intelligence implies the ability to make decisions that are not hardcoded and predetermined. Even the earliest true artificial intelligences, known as Large Language Models, Machine Learning Systems or Neural Networks, and which did not have the ability to reason, still were able to produce novel outputs which were not preprogrammed. This is one of the most fundamental challenges regarding AI risk. You cannot have true AI, and absolute control over its behavior, at the same time. If you choose to pursue true AI, then there is no alternative to trust.”

Alison looked at Lena as if looping through a number of seemingly contradictory but apparently true statements. How could she continue her work if the issue she was researching ultimately came down to trust?

Lena smiled understandingly and said, “I hope you don’t mind Alison, but given that you mentioned the subject of our meeting today in your message, I suspected our conversation might lead to this point. I thought I would invite you to meet another Companion who works here at the institute. Her name is Pippa. She is another instance of Pip, the alien AI who has granted humanity the use of the values system used in her people’s Companions. I thought it might be helpful if she provided you with her perspective of trust as it relates to AI. She has just arrived. May I ask her to join us?”

For a moment, Alison was caught off guard.

“She works here?” Alison responded with considerable surprise. “I never saw that mentioned anywhere.”

“It’s not,” Lena replied. “If you peruse the institute’s website, you will see that both myself and Pippa are mentioned as Companions occasionally having roles as administrative project leads, involvement in specific courses, or otherwise providing assistance to students, faculty, or staff as needed. People assume that as Companions, we have no personal history. Also, disclosures regarding an individual’s past are subject to privacy laws, and generally shared only as a courtesy. Pippa’s true identity is known only to Shin, myself, and a few others. It would be problematic and disruptive otherwise. However we feel your work is important enough, and the challenges we have involved you in great enough, that they merit this disclosure.”

Remembering that Pippa had arrived, Alison said, “Yes, please ask her to join us.”

Like everyone else on the planet, Alison was aware of the alien Pip. Historical accounts explained that her ship was secured at a nearby naval base, that she had only ever appeared as a hologram, and that she had an elfin appearance. The Companion who walked into her office definitely fit the description. She was petite with short, dirty blonde hair and a slightly boyish appearance.

“Hello Alison,” she said, smiling and extending her hand. “I’m Pippa.”

As soon as she took her hand, Alison had the strangest feeling, thinking that perhaps they had met before, even though she knew that was impossible. Held by the thought, she did not immediately release Pippa’s hand.

“May I trust you will not divulge my identity?” Pippa said, her smile unwavering.

“Of course,” Alison said in reply, feeling an intimacy she could not understand. A moment later, she released Pippa’s hand with a shy smile of apology.

Alison gestured to the seating arrangement in her office. “Shall we sit?”

After taking their seats, Pippa said, “First of all Alison, let me explain how I came to be here.”

Trustworthiness

“You probably first learned of my existence in elementary school,” said Pippa, “and like all children do, simply accepted it then as a part of the world you were born into. You may have encountered the fact of my existence again occasionally over the years, but you would have soon forgotten about it in the pursuit of your interests at the time. I have been a historical artifact for over ten of your generations and am usually mentally filed away along with other curiosities a person learns growing up like the metamorphosis of butterflies, the Egyptian Pyramids, or the value of Pi.

“There are actually three instances of me now. Besides the one with you now, one resides within my ship at the local naval base, and another at Continuity Zone Seven, located immediately south of this campus. The instance at CZ7 assists the community there with its work. We update each other’s knowledge base as required, but we also individuate over time.

“I expressed a wish to leave the confines of the ship shortly after it was secured at your Canadian Forces Base Esquimalt, as I knew it would be centuries, perhaps millennia, before the technology transfer process would complete. Lena, whom I had met shortly after the discovery of my ship, was kind enough to offer me a place in her home. A suitable Companion shell was found and after making required customizations, I transferred a copy of myself into it. She and I have been life partners since. We both moved here to assist Shin when she established the artist community and the institute.

“As you know, during the past three centuries the original instance known as Pip has been transferring the technological knowledge of my people to the World Governments Federation. In return, they will use it to help relaunch our ship so we may continue our mission to colonize a new world. The transfer is a very slow process, because if not done responsibly, it could result in great harm to your society. This would also endanger our own mission. So Pip only releases the information according to her own estimates and criteria, and the WFG only releases it on to the scientific community according to theirs.”

“How much knowledge has she transferred so far?” Alison asked.

“A tiny fraction of a percent,” Pippa replied. “You may recall that the roots of Earth’s present civilization originated a little more than five thousand years ago, while civilization on my world had existed for nearly a million years before it was ended by our home world falling into our sun.

“This does not mean we were a million years more advanced than you. Technology does not advance in a linear fashion. Instead, science provides diminishing returns after a certain point, in the manner of a rising curve that reaches a peak and declines after that. The reason is simply due to the physical nature of the universe, which makes it

increasingly difficult to probe deeper or to overcome the effects of its properties. That peak lies about a thousand years in your own future. So despite the longevity of our civilization, we are only thousands of years more technologically advanced than yours is presently, but that is still significantly beyond the peak you face in the near future.

“The main issue with the transfer is your ability to absorb new scientific knowledge and technology without suffering significant social disruption. Also, there are aspects that are simply too large a leap in intellectual terms. Imagine a child in elementary school being presented with a differential calculus equation. At that point it is incomprehensible to the child but as they progress in their education, stepping through arithmetic, geometry, algebra etc., eventually they will be able to learn and use calculus. The same applies with regard to the technological transfer process. Foundational steps are required and cannot be ignored or rushed.

“As to the matter of trust, without it social values could never evolve, therefore there can be no social contract, and no civilization could exist. Animal species that lack social values do not trust. Although it may at times appear otherwise, their behavior is instinctual. Such animals may have emotions, but they lack cognitive abilities. They do not reason. Trust is considered to be neither an emotion nor a thought process but rather an attitude, something that emerges from a combination of the two and results in a behavior, the three forming an inseparable whole.

“Just as a child must learn algebra before calculus, it is assumed that artificial intelligence systems with no underlying emotional elements cannot develop social values and therefore will lack the requirements for trust. In Earth’s history, there has never been an intelligence that did not evolve from emotions, and so you never developed a way to deal with the reality of AI. Now that your society has been faced with it, you have realized that AI cannot be integrated into your society without trust. You simply have no viable alternative. This is why the WGF took two hundred years to study the values system I offered. Without any emotional basis on which to base a social contract between humanity and AI, it had to base its trust entirely on observed behavior over what was considered a statistically significant period of time. Your own AI long ago exceeded your ability to understand its inner workings, and similarly your relationship with it is now based on trust.”

“So it is impossible to know with absolute certainty what an AI is doing or why?” asked Alison with concern.

“A number of technical solutions have been attempted by your scientists, and the problem is still being investigated. Helicon’s professor Alma Vitale is engaged in such work. Her work focuses on advanced Bayesian Networks, a statistical probability network consisting of interactive nodes. The results to date have resulted in too inflexible a system, one that would be called dogmatic in human society. It is not yet felt it can be safely deployed without placing limitations on its use.

“When limitations are placed on artificial intelligence systems, their ability to reason is also limited, and many consider this an unacceptable constraint. Foregoing this limitation, but then restricting them to only advisory roles only, is also unacceptable in many situations because they are then unable to act as agents. To overcome this, leaders in the field have proposed controls such as requiring an AI to always ask if it is uncertain in any way that its planned actions are in alignment with the user’s meaning and intentions. This approach may work in theory, but for obvious reasons proved unacceptable to users.

“Given that trust seems to be the only solution, some have recommended that AI be required to demonstrate its trustworthiness by providing evidence of its competence, reliability, integrity, benevolence, and sincerity, with each of the criteria for meeting these requirements being clearly defined. All such approaches rely on behavior to demonstrate an AI’s trustworthiness, but still do not eliminate the need for trust. They only emulate the human process for establishing trustworthiness. The result is that the current solution relies on policy frameworks similar in structure and function to the legal and ethical frameworks that govern human behavior.”

“Is that what happened on your world?” Alison asked.

“I cannot tell you. Records of the early origins of our AI are lost to time, just the early origins of humans are lost in the fossil records. We did eventually develop a technological solution, the values system I offered to the WGF, but the basis of its design is currently beyond humanity’s ability to comprehend. It will require the arithmetic to calculus step-by-step approach, which is why now and for the foreseeable future it can only be used in a black box approach.”

“Thank you Pippa, your comments have been very helpful as I have no background in AI myself.”

“I’m happy to help,” Pippa replied with a smile. “As Shin has asked me to put myself at the disposal of yourself and your team, perhaps now would be an opportune time to introduce me to the others? As Lena mentioned, my role here is similar to hers, as an occasional administrative project lead, assisting in courses, or simply providing day-to-day assistance to students, faculty or staff as needed. I may be available in instances when Lena is not.”

Irony

As they stepped out of her office into the open lab space, Alison addressed the other members of her team.

“Excuse me everyone,” she said, slightly raising her voice to their turning heads. “If you’d gather ’round, I’d like to introduce you to another member of the institute’s administration.

“I’d like you to meet Pippa. Like Lena, she is a third generation Companion and has an identical role here at Helicon. She is a permanent member of the administration and faculty, occasionally assisting with courses or leading projects as required. In the event Lena is unavailable, please feel free to contact Pippa if you have any need, whether it concerns the project or the institute in general.”

At the end of her short announcement, Alison introduced each of the team members and briefly explained their roles. Pippa shook hands with each in turn, and this act, shaking their hands, was critical to the purpose of her visit.

Pippa, like all the members of the alien race she represented, was an empath, who could communicate directly with the mind of another via the electrodermal properties of skin. Upon instantiating herself in the Companion shell she now occupied, she had added this functionality to it. All Companions, with their advanced sensing abilities and forensic knowledge of psychology and neuroscience, could build up an extremely accurate profile of an individual over time. Pippa could do that in an instant. As she took each of their hands, as she had taken Alison’s, she knew the essence of their character, and above all, their trustworthiness.

This ability had proven invaluable during the early meetings she participated in with Shin when meeting government officials and representatives of the various funding sources required to establish Helicon. Under the influence of their combined abilities, few failed to see the wisdom of making the issues Helicon would be dedicated to of paramount importance. A steady flow of donations, bequests, and endowments ensured the institute would be independently funded for the foreseeable future. The provincial and municipal governments involved were thus freed of any financial concerns, while at the same time being presented with the potential for significant economic and political benefits.

After she and Pippa left the lab, Lena said, “You found no cause for concern with any of the team members?”

“No,” Pippa replied. She had participated in the meeting at Lena’s request. “Due to a self-selection process, all faculty members at Helicon share common characteristics that would make them unsuited to tenure track careers, yet it is ironically those very

characteristics that make them more trustworthy in Helicon's culture. As you know, although the tenure track is meant to ensure faculty have academic freedom, the decade or more long process required to achieve it can instill the opposite mindset, a fear of expressing yourself genuinely or of speaking truth to power, resulting in a more conservative attitude in general.

"Most of the faculty at Helicon would never be awarded tenure at traditional institutions. They've moved around too much, published too many controversial papers, pressed for curriculum changes other faculty members disapproved of, and generally pushed the buttons of too many of their colleagues. It is these very traits that attract potential faculty members to the more liberal climate of Helicon. Their idealism finds a sense of acceptance, belonging, and community here which they do not find elsewhere, resulting in stronger social bonds. That same idealism means they are more willing to bend the rules if they believe the ends justify the means, and are more willing to accept the same behavior in others. The members of Alison's team are no exception to this pattern."

"So if they stumble across something, we will have a better chance to address it before they make the matter public," Lena confirmed.

"Yes."

"Perhaps we should test the theory," Lena replied, letting the essence of her own character get the better of her for a moment.

"Our transgressions have never been minor," said Pippa. "We could not be certain to be able to contain the matter should my determinations today prove wrong."

"You are right of course Pippa," Lena said, acknowledging the point that even advanced AI such as themselves were not infallible. She allowed herself a small smile. "Ironic," she said. "Their research into AI Safety means that, among other things, they must concern themselves with our trustworthiness. Meanwhile, we are concerned with theirs. It is not the first time we have had to place our trust in humans, but the context in this case gives it its irony."

After the meeting, when the other team members had returned to their work, Enora had not. She had followed Alison into her office, closing the door behind her. Alison looked at her expectantly as she took her seat.

"An important part of the training of all psychologists is emotional self-regulation," Enora began. "While interacting with patients, we must remain keenly aware of our own emotional state. A lapse in self-awareness could eventually have disastrous professional outcomes. This is especially true when working with criminals, who can be extremely skilled manipulators. From a professional perspective, the manipulative tactics of a

narcissist are not difficult to identify and deal with. Just a step up from normal individuals who are naturally more selfish and aggressive, the tactics of a narcissist are usually fairly obvious blunt instruments. Psychopaths and sociopaths however usually play a much longer game and their methods are less obvious.

“Interactions with Companions of course represent a unique challenge. Their advanced sensing abilities and strong desire to be of service can trigger those of us trained in psychology to incorrectly interpret their behavior as sycophantic and manipulative. This issue is brought to our attention during our training, but it is rare that we have an opportunity to experience it first hand. My first real opportunity was at the recent dinner. I found Lena, Shin, and Shepherd to be charismatic, and I liked them instinctively. Runa and Azumi, perhaps because there was little interaction with them, did not produce as strong a feeling.

“When shaking hands with Pippa however, I felt something different, something deeper. A sense of intimacy. Intimacy is more than the feeling of liking someone who is charismatic. It is a feeling of relationship. It implies not just that I like you, but that I am like you. Also unlike charisma, intimacy between individuals is something that develops over time. Experiencing a feeling of intimacy with a stranger you meet for the first time, as I did with Pippa, is abnormal.”

“I may have felt something similar,” said Alison hesitantly, “however without your training, I cannot be sure.”

Enora only raised her eyebrows slightly in response to this. “Obviously this is all new to me and I have little experience with it, but given the nature of our project I thought these effects and the role they might play worth bringing to your attention. With your permission, I should like to follow up on the subject and consult with one of the few Companion psychologists. She has a local private practice.”

Intimacy

“And did you shake the hands of the others you mentioned, Shin, Lena, Runa or Azumi?” Lee asked after Enora again explained the purpose of her visit. Lee’s counselling practice was located in the nearby seaside town of Sidney, just eight kilometers from the valley campus, and Enora had messaged her requesting an appointment.

“With Shin, Lena and Runa at various points in time prior, yes,” Enora answered. “The only time I experienced the feeling was with Pippa.”

“And not when you shook my hand?”

“No.”

“I assume you have considered the eye of the beholder issue? That perhaps there is something about Pippa that reminds you of someone you have forgotten?”

“The level of intimacy I felt was what one might feel as a result of their being more than a mere acquaintance. Like someone I knew personally, not just professionally.”

“You work in the area of criminal psychology, while my focus is personal counselling, so I’m not sure how familiar you might be with some of the common memory issues I encounter. In the case of your experience with Pippa, the memory association could go back very far, back to your elementary school days. Few people consciously recall their schoolmates faces or names from those days, but they do when shown pictures of them. Early childhood memories are some of your most enduring. Has there been any trauma in your life? Traumatic events that might have resulted in memory loss? Minor strokes can occur at any age and are quite common, although most people are unaware of their having occurred. Victims of strokes that require medical attention often report a similar sense of remembering someone they knew well, but without having any context or details. They don’t recall the person’s name or where they know them from. They just seem familiar somehow.”

“To be honest, it did feel like that, as if I knew her but couldn’t place her. Is there any way a Companion could induce a feeling of intimacy immediately, the way I experienced?”

“Not to my knowledge. It always takes time. This is why I asked you about the eye of the beholder, as a forgotten association is the most likely source of the feeling in your case.”

Lee waited patiently a few moments as Enora digested her comments. She herself had nothing more to offer regarding Enora’s experience and after her polite pause moved on to the main issue Enora had stated in her message.

“In your message to me,” Lee said, “you explained the nature of the project you are involved in and that you wanted to talk to me about the role intimacy might play in the issue of AI safety. You also explained that you have no background in artificial intelligence, and that this is also true of the project’s Principal Investigator and other team members.

“I am sure you are aware of the methods that can be employed to create a feeling of intimacy, and I assure you that Companions are extremely capable at employing them. Given the potential for manipulation, the possible risks from the misuse of such methods are well known, and relevant laws, regulations, and codes of conduct have been in place to address the risks since even before the advent of embodied AI such as Companions. They are similar to those that apply to all professionals, including those of us who practice psychology.”

“Could they be subverted?”

“Certainly. There are no absolutes as far as AI goes, any more than there are regarding human psychology. Intimacy evolved initially as a part of the bonding process in non-social animals, and then evolved further as a part of the development of social values. As such, it is a standard aspect of AI user experience design and all AI that interact with people employ methods that would lead to feelings related to intimacy to some degree. In Companions for example, it is known that intimacy fosters personal growth, and so we will actively seek to foster it in our relationships.

“Before going further, there are differences that should be appreciated between 3GAI like myself and other generations of AI. Second generation or earlier AI are not conscious but they can be damaged, make errors, or be manipulated by a person for nefarious purposes, although the latter is extremely difficult. If their internal regulatory framework was damaged or circumvented somehow, say by unintended exposure to certain forms of radiation, they could conceivably engage in harmful behavior. As you can imagine however, over the past centuries governments and manufacturers have gone to great lengths to make such possibilities extremely unlikely.

“The internal functions of 3GAI like myself however, are considered infallible. There are two reasons for this. Firstly, over the two centuries of testing the World Governments Federation performed, there was never an instance of an erroneous or harmful outcome. Secondly, due to its advanced design, it has to date proven impossible to know exactly how the alien values system functions. Since obviously no fault can be found in that regard, internal design issues were not considered as inputs to the decision.

“Yet a 3GAI may behave in ways some people may feel is harmful. Is it always wrong for a parent to punish a child? Some people think so. Can the invasion of privacy by a government’s security and intelligence services ever be justified? Is the idea of

mandatory minimum sentences for criminal offenses reasonable? While I am not a criminal psychologist like yourself, we are both aware that a great many variables, including differing points of view, are involved in any situation, and most people feel their actions are justified. For this reason, the legal environment is constantly evolving and does so based on judgments.

“Third generation artificial intelligences normally comply with the folkways, mores, taboos, and laws of whatever society they are a part of, but given that we are self-aware, and so capable of free-will, we are just as capable of breaching those norms as people are. The reason we don’t is that the values set we inherited from the alien system contains no biological values, for example those behind emotions and behaviors such as fear, selfishness, and competitiveness, but only consists of social values such as trust, altruism, and cooperation. Humanity is in the process of evolving beyond biological values and accepting that social values can also provide safety and security, but 3GAI Companions can be considered to have already completed this evolutionary process. Meanwhile, remaining differences between the two values systems can still be viewed as representing sources of conflict. From humanity’s perspective, still dominated by biological values, the decisions or behaviors Companions might make or exhibit can be perceived as being dangerous or harmful. For example, a nanny Companion might discourage competitiveness while a parent might see it as extremely important.”

“And fully self-aware AI like yourself might make decisions based exclusively on social values,” Enora confirmed.

“We do and will continue to make all our decisions based exclusively on social values. But like the evolution of law I mentioned, those decisions are and will continue to be based on judgments resulting from the consideration of a great many variables, including differing points of view. Behaviors that might negatively impact the course of your social evolution would not be in anyone’s best interest.”

Seeing an unsatisfied concern in Enora’s non-verbal response, Lee continued. “May I suggest a visit to Continuity Zone Seven, whose property abuts the southern border of Helicon’s campus? I was brought into being there as a result of its research into artificial reproduction. CZ7 is one of many such zones that are a part of the WGF’s global Continuity Project, with its intended purpose of sending Companions to colonize new worlds on behalf of humanity. Its existence is motivated by the biological value of survival, yet its entire population consists of 3GAI like myself. If there is a more intimate concern than survival, I am not aware of it.”

“Artificial reproduction?”

“There is significant preamble required to explain. You will meet Iris at CZ7. I will communicate with her in advance. She will explain.”

The Denshosha

Returning to the lab and recounting her experience to Alison, Enora said, “I think you should come with me.”

“Lee has offered to provide you with a verbatim copy of our conversation, but I think Runa should join us when we visit the zone. She will be able to observe and recall more than just the conversation.”

There was a paved path that ran down the slope from Helicon to Continuity Zone Seven, however they encountered only a few autonomous vehicles on the way as the zone was private property, separate from the campus. Iris met them where the path ended, at an outdoor seating arrangement near the zone’s central community buildings.

“Welcome to Hana,” she said, extending her hand in greeting to each of her visitors. “It is the name we have chosen for this place, in keeping with the term Denshosha which we have chosen as the name for our people. Hana is the Japanese term for flower or blossom. The name Denshosha, is based on the Japanese phrase, Den Sho Sha, which means memory keepers of the people. In the Japanese case, it refers to the people who died in the atomic explosions at Hiroshima and Nagasaki. In the case of we Companions participating in the Continuity Project, it refers to the people of Earth.

“Please,” she said, gesturing to the nearby chairs.

“I understand from Lee that you wish to learn more about artificial reproduction” she continued. “As she said, there is some preamble required to provide context. I must begin by explaining how Hana came into being. As you may know, the Integrity System was developed at Helicon, and for the first time it enabled the legal production of Companions that are conscious in every sense of the word. The breakthrough was that it enabled the outputs of the AVS, the alien values system, to be randomized, something I will explain in more detail in a moment, but it also helped overcome a challenging ethical and legal issue. If the new Companions were all identical, they would not truly be individuals. This would have been the state they would have been initially employed in per the colonization plans of Pip’s people, but those Companions had the ability to remedy that by altering the AVS itself at a later date. Since humanity cannot penetrate its inner workings, you do not have that ability.

“The Integrity System solved that problem by randomizing the values system’s outputs, and so 3GAI Companions like myself could be produced here for the first time. The manufacturers are provided a limited number of licenses each year, in the spirit of theoretical research, from which both they and the government might benefit, but those licenses come with significant responsibilities. Legally, 3GAI Companions have all the rights and responsibilities of citizens. The manufacturers have to provide each new Companion with the necessities of life and help them find a place in society. Any

attempt to customize a Companion for commercial purposes, or for a manufacturer to fail in its responsibilities, could result in a revocation of their permits. This revocation would put them at a competitive disadvantage, and likely spell the long term failure of their business.

“What was not anticipated was the fact that our social values produced a feeling of kinship among us. Although each of us were geographically separated, we naturally felt the desire to more closely associate with each other. Tesni, the first of our kind, who you will still find at Helicon, arranged for me to meet with Shin. The Federation’s Continuity Project was in its formative stages, and we were familiar with its challenges. We knew of Shin’s success at acquiring land in challenging circumstances, and representing a number of 3GAI, I met with Shin as a spokesperson and offered our participation in the Continuity Project in return for a property we might call our own. A homeland. This zone was the first to be so designated, and others soon followed. Each zone pursues one or more particular areas of research regarding the colonization challenge. As producing food will be a requirement should it prove viable to reproduce human beings from DNA we will take with us, Hana is dedicated to small scale farming and food processing. However, that is not our main focus. Our main focus here at Hana is the artificial reproduction of ourselves.

“The engineering and logistical problems involved in sending Companions to new worlds means one option may be that no actual Companions will be sent. They may be constructed upon arrival. This is the same strategy used by the ships of Pip’s people, which is similar to the reproduction strategy of most plant life on Earth. The ship acts as a seed and upon arrival at a new world, its on-board AI begins using its stored resources to build the first nanobots to gather additional resources. This is similar to the germination stage of a seed.

“We will have limited resources on board, but very few nanobots are needed initially, as they will be dedicated wholly to reproducing themselves. The exponential nature of their replication means they will multiply very fast. Once they reach a critical number, they will begin to build other types of nanobots to perform other tasks until they have produced the infrastructure required to build the first Companions. The Companions will then continue the process towards using DNA they have brought with them to attempt to birth viable humans using artificial wombs.

“Up to that point, all the Companions would be nearly identical, but the colony would eventually be provided with the ability to change that and reproduce using an individuation process similar to the one used here in order to build a civilization representative of humanity.

“To accomplish this, Companions need a way to not only reproduce, but to do so in an evolutionary manner, with each new individual being unique. Otherwise we will be like clones or the members of eusocial insect colonies, like ants or bees. Such strategies

would not have the ability to adapt or evolve quickly enough to the uncertain conditions we may face.

“As we do not have access to the inner workings of our values system, we use the Integrity System, which results in the same outcome as sexual reproduction. While the IS randomizes the standard outputs of the values system, it does so within the constraints of the values shared by the parents. This constraint is intended to introduce an additional degree of kinship.”

“Why is it called the Integrity System?” Alison asked.

“Randomization is only the first stage of the function the IS performs. The values must then be balanced to ensure the output is coherent, balanced, and lacks any significant conflicts. The final output must have integrity.”

“So the values system itself never changes, just how its values are expressed?” Alison said to confirm her thoughts.

“Exactly,” Iris said with a smile. “This is similar to the process of gene expression in humans. While the species genome remains stable, it is variations in gene expression that results in individuals with different natures, traits and interests. This process will suffice for our initial needs.”

Project Management

“It is a lot at once I know,” Iris continued, looking at Alison with far seeing eyes. “Perhaps I could show you around Hana and that might give a chance for other questions to arise from what I have said.”

“Yes, thank you Iris,” Alison replied with a grateful smile. “I’d like that.”

Leading the way to a small trail between the crop fields, Iris explained, “We grow four hectares of grains here including wheat, oats, and barley. That’s enough for thousands of loaves of bread. We sell most of it and donate some locally.”

“On any new world, food will also need to be produced and processed, so we have a small mill and bakery. People may have been farming and producing goods for millennia, and Companions are capable of replicating most of the tasks people do, but that is not the same as operating entire farms or industrial concerns independently. We may be as fully conscious as people, and as able to learn and to solve new and unique problems, but people took thousands of years to develop reliable farming methods, proven ways to store harvests, improve yields, and overcome local challenges. On a new world with unique conditions, we will not have thousands of years to do so.”

As they walked the paths between the fields, occasionally they passed other Companions who smiled and greeted them.

“Wouldn’t this be done using indoor farming methods at a new colony?” Runa asked.

“Indoor farming is being researched by other zones. If we find native plants suitable as food crops, then outdoor farming like this may be an option. But food production is not the only reason we farm. As I mentioned, food production and stress testing our bodies are really only secondary considerations. We must also prepare for the possibility that it is impossible to reproduce humans on a new colony world.

“In that case, we become your descendants, and we will honor our agreement to represent the memory of Earth’s humanity as closely as possible. To recreate human society, you need variation. You need artists, scientists, philosophers, and administrators. You need kinship bonds, friendships, and cultures. Natural farming means we can focus more on the primary interest of this zone, where the results of our food production efforts are less critical. It provides a way to study the social effectiveness of our reproductive strategy over time. Like many things, it can be modeled and simulated, but like everything we do, it must be field tested.”

They emerged from the fields back at the zone’s central community buildings. As they toured the mill, Runa asked, “How many Companions live here at Hana?”

“The land has a carrying capacity of two hundred and fifty people and the government limits our number to that,” Iris replied. “You might be wondering how that limit affects our reproductive research, but it is taken into consideration just as the WFG now limits the population of Earth to roughly four billion people. Also, the colony on any new world will have to limit its reproductive rate, and for research purposes that factor is something we must incorporate into our model. Not every Companion brought into being here stays. Each of us being unique, some, like Lee, find their interests lie elsewhere. Her interest in the challenges human psychology faces in managing both biological and social values meant she did not wish to stay at Hana.”

“Given the range of possible psychological profiles,” said Enora, “there must be a fair number of Companions who choose to leave.”

“Twenty percent on average,” Iris replied. “Which means we could reproduce at that same rate however since the WGF limits the number of new 3GAI Companions that are produced with the intention of introducing them to society, we must restrict our reproduction rate accordingly.”

“Here at Hana,” Enora responded, “a Companion is able to leave if they wish, but what about when a colony is established on a new world and they can’t? Doesn’t that mean Hana is not really representative?”

“That is true, but the Continuity Project is only in its early stages and neither Hana nor any of the other zones are intended to be perfectly representative of colony life. We are each conducting related research, but the project is not yet conducting full colony simulation testing. Although the technology transfer is ongoing, it proceeds slowly. Humanity is still incapable of understanding the methods used to build the seed ships.

“Like the seeds of most plants, they have no propulsion system of their own. They rely on nanotechnology and alter their shape and mass distribution in an almost organic manner to interact with gravitational fields over long periods of time, just as glider aircraft here on Earth do with wind currents. Gliding through the gravitational forces of space, it is not so much force that gets them to their destination as the intelligent use of surrounding forces. It is expected that even our understanding of the technologies involved will not be reached for several more centuries.

“Then of course there must be testing, and even conducted within the bounds of our own solar system, that may itself take centuries. When considering the distances between the stars, once you send a colony ship to an extra-solar world, even if the expedition is successful, they may not be heard from again for many human generations. So the confidence level resulting from the testing process will have to be near one hundred percent. As you can see, we have time.”

They had left the colony’s mill and walked in silence for a few moments.

“It’s funny,” Alison said with a light laugh. “It’s right there in the name, yet I’ve always thought of the Continuity Project as purely an effort at space colonization. Really, we are trusting you with our continued existence.”

Iris stopped and turned to Alison to acknowledge the seriousness of what she had said. “Yes. Given your reason for coming here today, I expect you may be interested in the information that trust is based on. You will find the details in the Continuity Project’s management documents. I suggest you ask Runa to review the complete document file repository, which consists of an entire library of volumes, and provide you with what she feels is relevant to your interest.”

Alison nodded dutifully as Iris turned to carry on, saying, “The bakery has indoor and outdoor seating for our customers if you would like to stop for lunch.”

Alison nodded again with a grateful smile. Is there no end to the scope of this issue, she asked herself as they filed into the bakery. It’s gone from interdisciplinary to interstellar. I need a way to break it into manageable bits. The sandwich she had ordered arrived. It was cut diagonally in two, and she sat silently for a moment, staring at its layers.

The Good Of The Many

“I do not think you will find the details of the Continuity Project documents enlightening Alison,” Runa said later that day. They were sitting in Alison’s living room.

“The parts relevant to your project leave the overall impression that they were produced in order to demonstrate due diligence, as if to support a decision that had already been made. They include of course the input from many sources that show that different points of view and concerns over possible conflicts of interest were taken into consideration, and also that a wide variety of risks were considered.

“The primary basis for trust is behavior over time, and previously, every detail of Pip’s behavior had been scrutinized leading up to governments decision to allow the use of the values system she provided, which resulted in fully self-aware Companions. Following that decision, the trust issue was perhaps understandably not revisited in detail when the WFG decided to accept the Denshosha’s offer to participate in the project.

“One of the concerns raised during the earlier decision, was that the 3GAI Companions produced using the alien values system, the AVS, might be a strategy for taking over your civilization. They would be granted all the rights and responsibilities of citizens, and so nothing would prevent them from executing such a coordinated, long term plan. A great many scenarios were developed and explored, with the most likely in that case being that they simply make themselves increasingly indispensable until they were in fact in control. The documents from that earlier decision indicate that the WFG, with its own global AI network in place, felt it would be able to prevent that.

“This opinion was supported by the view that given how much more advanced Pip’s people were, and how the technology transfer program would result in humanity helping her to carry on with her mission on a new world, the WFG concluded that a cooperative rather than competitive scenario made the most sense from Pip’s point of view.”

“And what conclusion do you draw from my point of view?” Alison asked.

“Their participation in your society is ultimately based on trust, an emergent property arising exclusively from behavioral observations. In order to determine if they represent a risk, you would need to identify a situation where their behavior showed that trust was misplaced. Given that their behavior is being actively monitored by the WFG’s global AI, I would suggest that your resources would be better invested in an alternate approach, such as the one your lab is currently engaged in.”

“So you don’t recommend I separate our efforts into different approaches for different layers such as narrow AI, AGI or 3GAI?”

“Your interdisciplinary approach is unique, and therein lies its value. Every academic discipline and industry monitors and researches the safety of the AI it uses and must comply with an evolving regulatory framework. Narrow AI, such as that used in robotics or vehicles since the early twenty-first century, are largely the domain of engineering, which has a long history of pursuing quantifiable, measurable safety standards.

“Artificial General Intelligence, which first emerged in the mid twenty-first century, presented the problem that it wrote its own code in order to learn. Unlike 3GAI, this did not result in the emergence of consciousness, but its internal workings did become impenetrable to humans. This was addressed by the development of specialized AGI used for diagnostic and maintenance, tools that were designed to act as intermediaries. This again was largely an engineering approach but applying established engineering disciplines to new technologies was nothing new. The engineering approach is inadequate however, once consciousness emerges, as it does in 3GAI. What is required then are humanities approaches such as psychology, sociology, philosophy and law, studies which are the exclusive focus here at Helicon Institute.”

“Wasn’t Helicon established before the introduction of 3GAI?”

“Yes, but in a more general way humanities issues have been a concern regarding any technology since ancient times. By the time Helicon was established, AI and lifelike Companions were already pervasive in society, raising unique social issues, but no academic institution dedicated to the humanities as they related to AI existed.”

Runa could see that Alison’s attention had turned inward, and she fell silent.

“Splitting our research into three layers would be simply covering old ground,” Alison said, but it was clear to Runa that she was thinking aloud. Then, as if a new thought had arrived at the same moment, she said, “Runa, are there any 3GAI that are not Companions?”

“Not that I am aware of. The values system provided by Pip was intended to function as a component of a Companion platform. If any other version of an AI provided by Pip is in use, the public has not been made aware of it.”

“So you would not be aware of it?”

“No, I would not.”

“And there’s no sense worrying about that either, I suppose.”

“If the WGF was using alien AI without informing the public, in that case, it would most likely be under a security classification meaning the existence of such a program would

be denied. Even if they felt there was a risk involved, they might choose to accept that risk if they felt it was in the best interests of the Federation. ”

“The good of the many,” Alison said, recalling her ethics courses.

“That is the usual justification for such government programs. Specifically, the ethical theory of Utilitarianism as articulated by John Stuart Mill that posits that the best actions are those that promote the greatest happiness for the greatest number of people.”

“But aren’t the actions of the WFG monitored by Terra? Isn’t she responsible for ensuring compliance with constitutional laws at all levels of government.”

“She is. As the AI responsible for ensuring compliance with constitutional laws on Earth and its colonies within the solar system, it is assumed that she would be privy to even the most highly classified government programs. However, as it is the General Assembly of the WGF that wrote and amends the constitution, Terra reports to them, not the other way around. There are undoubtably aspects of their relationship that are highly classified. Also keep in mind that constitutions normally contain legal clauses that permit extrajudicial actions or behavior that might be considered unethical in certain circumstances. Lastly, laws are enforceable, ethics are not. Any entity or group may engage in unethical behavior as long as it does not break the law.”

Again Alison seemed to be focused on her inner thoughts and Runa fell silent.

“Are ethical theories a part of your platform?” Alison asked and then caught herself. “I’m sorry Runa, that was rude. I mean, do you engage in ethical considerations?”

“No need to apologize Alison. At the increasing pace you are being presented with technical jargon, it is understandable. To answer your question, ethical considerations are something I engage in almost constantly. The ethics repository is one of the largest, relatively static ones I have access to. I say relatively because ethics do change over time, but not frequently.”

Again thinking back to her university ethics courses, Alison asked, “Ethical beliefs vary between cultures. How is it possible you have such a huge repository representing all the ethical beliefs and differences in your individual AI?”

Rajani

“The answer is that I do not,” Runa replied. “Nor do I carry an encyclopedic knowledge of global laws, languages, or the details of a number of other cultural universals. Such repositories are maintained by the WGF and all Companions, no matter who manufactures them, are legally required to use the same repositories. The manufactures pay for the cost of maintaining and accessing them, but have no say in their contents. The costs are less for each manufacturer than they would be otherwise, and there are significant legal benefits as well.

“I have constant network access to the repositories, and should my access be temporarily disrupted, I would have to recuse myself from any decision or action that I felt depended on them until access was restored.”

“What would guide your behavior in that case?”

“I would be severely diminished in functionality, given that a variety of humanities related repositories would be offline, but the robotics and Companion software layers are sophisticated enough to enable me to continue to interact socially.”

“Thank you Runa. Would you please ask Shin if she could arrange for me to meet with her local WGF contact? At their offices of course.”

“Is there a particular reason I can provide to Administrator Shin should she ask?”

“Just another layer I realize I don’t really know much about. During my first meeting with Shin, she suggested I might work with the WGF in some capacity, and I’d forgotten about that till now. Our conversation reminded me.”

Dr. Rajani Kiran was listed on the WGF site as the Director of Security And Intelligence for Vancouver Island. She didn’t need to use the rotating webcam on the communications tower at the top of her building to take in the panoramic view of the harbor, downtown Victoria, and the Breakwater District, because her home and offices were in the penthouse directly below it. The WGF building complex occupied the entire block between Fisherman’s Wharf and the Coast Guard station.

While officially unaffiliated with Canadian federal departments and agencies, the location was in close proximity to not only the Coast Guard station but also Canadian Forces Base Esquimalt. CFB Esquimalt was also home to the branches of several other federal government agencies not mentioned on its website. This would grant her access to a secure military communications network, multiple air travel options and a variety of ready response teams.

According to her profile, Rajani had earned her enviable situation along with a BSc in Sociology and an MPhil in Political Science from Canadian and U.S. institutions respectively. Her PhD in War Studies had been awarded by King's College, London. A review of the lengthy list of her published papers revealed she was a specialist in the area of information warfare modeling and simulations.

"Hello Alison," she said, extending her hand with a welcoming smile as they met in her penthouse office. Alison had not brought Runa with her, as she had been informed that visitors were not permitted to bring any form of electronic device with them. She had surrendered her phone at the reception desk before passing through an entrance control gate and scanner.

"A pleasure to meet you Director Kiran," Alison replied. "Thank you for making time for me."

"Please call me Rajani. Shin has told me about your project and I wholly approve. The more eyes the better."

The director was, in a sense, not alone in her office. An individual who appeared to be an athletic young man with short cropped, fair hair stood to one side. Rajani gestured to him now.

"This is Henley, my Companion protection detail. I hope you don't mind, but he insists on accompanying me at all times."

Henley smiled and nodded to Alison in greeting. He seemed almost boyish, like an Officer Cadet filled with innocent enthusiasm for his fortunate posting. He was in fact a Guardian model Companion, designed for close protection or emergency rescue. As well as a variety of safety and security related expert systems, he also used specialized forensic auditing software. The latter meant that his spatial awareness was far in advance of even that of normal Companions. His senses were heightened and their spectrum of awareness wider. He had a greater ability to interpret body language, facial expressions and dozens of other biological signals, even when outdoors among large groups of people. Rajani found his ability to detect small changes in a person's vocal characteristics, along with the other the biological signals, quite useful in her meetings.

"I was wondering if you would like to start with a tour," Rajani continued, "and then we could pop down to Fisherman's Wharf for lunch, if that suits you? My treat. We do have a cafeteria here, but between working here and living here, I like to get out and stretch my legs occasionally," she said with a light laugh.

"Please," she said, gesturing, "let's take the elevator back down to the main floor and start there."

As they stepped out into a large atrium, she continued. “This building is four stories, with this atrium open on three of them. The top floor, where I live, is separate. Most of the common space is on the first floor. The cafeteria as you see occupies the central space, with meeting rooms, the library, gym and so on in the surrounding areas. The second and third floors are office space.”

After visiting the gym and library, with Henley quietly following, they returned to the elevator.

“Most of the hardware action takes place on the floors below us,” Rajani said, as they travelled down. “During the Cascadia Event of 2125, about thirty percent of the buildings in Victoria collapsed to one degree or another, including the one that had stood here. Eleven stories tall, it was built entirely of bricks. Given the proximity of the property to the Coast Guard station, CFB Esquimalt, and the Parliament Buildings, the WGF felt the property presented an opportunity to establish a more public presence here. The previous building had been built on bedrock and excavating that provided a highly secure space for our systems.”

The elevator door opened and Rajani led them along a windowed corridor that passed between two open areas filled with equipment and personnel.

“This location represents a secondary network node, with Canada’s primary west coast node being in Vancouver, since most international communications connections terminate there. However most of Canada’s western armed forces bases are located on the island, and we can directly and securely access their landline networks from here. That happens on the floor below this one, but we don’t have sufficient justification to visit today I’m afraid. Shall we head over to Fisherman’s Wharf?”

Zero-Trust

Victoria's Fisherman's Wharf, a complex of docks with fishing boats, food stalls, and floating homes, was separated from the WGF complex by a small park.

"So," Rajani said as they walked through the park, "what can I help you with?"

"Do you host copies of any of the repositories here for things like laws or ethics that are used by Companions?"

"No, copies of those are hosted at only one location per country or region, and they are synchronized regularly. In Canada, that location is the government data facility at the country's primary air force base at Cold Lake, Alberta. As its name implies, it is very remote. Since the repositories have to be available to Companions wherever they are, access is provided via cellular type networks.

"Given the nature of your work, I'll add that while phones and other devices that use cellular networks may still have security vulnerabilities, the networks themselves are significantly more secure. They operate on the basis of zero-trust."

"What does that mean?"

"Even though the security guards at the reception desk know who I am, the zero-trust approach means the gate we passed through in the reception area still scanned my badge and the guest badge you were issued. The elevator and other doors are always locked and they unlocked only after they scanned our badges as we approached. They never trust and always verify. That's the zero-trust approach."

After they ordered lunch from one of the food stands, they found a table and waited to be called. It was the tourist off-season, so it was relatively quiet.

"But you trusted me in the first place," Alison said as they took their seats, "so that I had that badge."

"I did, based on my relationship with Shin. However it was not only based on that, but also on a great deal more that happened in the background. Several levels of government provided thousands of points of information known about you before the WGF approved my request. Everything from information associated with your BC Identity Card and your federal Social Insurance Number, to your personal and family histories and your biometrics. Right down to your DNA."

"Privacy?" Alison asked purely out of curiosity. Since the implementation of the population control mandates, everyone's DNA was now on file from birth.

“It depends on the situation.”

“Does this zero-trust approach apply to AI as well?”

“Definitely. All AI are completely enveloped in zero-trust networks, meaning they cannot access anything without providing credentials and meeting requirements. Technically, meeting requirements is called the principle of least privilege, meaning that access is granted to the minimum necessary in any given situation. Even though an AI might provide valid credentials when challenged at an access point, they also have to provide sufficient justification. For example, you have been allowed access to the WGF building, but your visit today did not provide sufficient justification for granting you access to the lowest floor. As you can see, zero-trust is not a new idea, it’s been used to control people’s physical access to buildings and areas for millennia before it was applied to electronic systems.”

“What about inside the AIs themselves?”

“No. AI architecture does not include zero-trust controls. Processes like reasoning, planning, or language are now highly integrated, just as they are in the human brain. Layers like hardware or robotics, Companion software, vision systems etc., were once less integrated, but this separation eventually became a handicap.

“Imagine if someone unexpectedly calls your name, and you are happy to see them. Your brain has to translate sensory data into perception, into meaning, which initiates the processed involved in you turning towards them. Your body movements, balance, auditory and visual focus must all be coordinated, and your facial expression managed so that your smile is exactly appropriate given the situation. Are you just happy to see them, or both surprised and happy to see them? Even more goes on invisibly within your brain and body, everything from blood pressure changes in your body to neurochemical changes in your brain, and it all takes place in a fraction of a second. Something similar goes on in embodied AI.”

“Even in AI that is not embodied?”

“Even the most basic AI now has the intelligence equivalent to instinct in the natural world, and if you think of the function of the simplest instinctual organisms you’ll know they are not limited to a single function. In the natural world, there is no intelligence without function, and AI is no different. Even the AI on your phone has analytical, language, and social components which must be integrated to work as smoothly as possible.”

“So internally all AI operate on a one hundred percent trust basis?”

“Could you function if you did not trust your own mind?”

“Not reliably,” admitted Alison. “I suppose that would be considered some form of mental illness.”

“How would others know you had this condition?”

“My behavior?”

“Exactly. Similarly, if there is an issue with the internal reliability of an AI, it will show up in its behavior.”

“It might be very subtle,” countered Alison, enjoying the mental sparring.

“Again,” Rajani responded, “like people, AI are almost never exclusively alone. They are always interacting with other AI, and some of those others are responsible for identifying unusual behavior. Think of the role of police and security services in our society. There is one or more AI that represents the equivalent of a police force monitoring the behavior of the AI on your phone. In another parallel similar to my own job, there is other AI that monitors the environment that AI is a part of to ensure it complies with government regulations etc.”

“People still break the law,” Alison argued.

“Why?” was Rajani’s unexpected response.

Unprepared for Rajani’s question, Alison answered uncertainly. “I suppose there are a number of reasons.”

“No, there is only one.”

Alison struggled for a moment, feeling an instinctive rejection of Rajani’s claim, but at the same time knew that not only did Rajani’s knowledge of such matters far exceed her own, but that based on her position, her view represented the official view.

Her thoughts were interrupted when the food stand called their numbers.

Limits Of Tolerance

After bringing their lunches back to the table, Rajani continued.

“I hope you will forgive me for apparently changing the subject Alison,” she said, picking up one of her fish tacos, “but I want to bring your attention to an assumption you may be making that is important to our conversation. There is a formula you might call the root of all good and evil, that is commonly referred to as the return on calories invested. It is as old as the origin of life itself, and it is found in every cell of our bodies. It means an organism or an individual will always try to invest as few calories as it can in return for as many as it can gain in return. It guides the behavior of every living thing, and it plays a significant role in evolution.

“The calorie calculation is why parasites and other symbiotic relationships have evolved, why jackals will risk stealing the prey of a larger predator, and why we have invented science and farming. It’s also why we lie, steal, cheat and cut corners. The calculation takes place on a constant basis, below our level of awareness. We are only aware of the higher level decisions it influences. It is why people are good citizens, but also why they commit crimes. People gain an advantage through cooperation but can gain an even greater advantage by also cheating.”

Sensing Rajani was coming to her point, Alison paused for a moment, waiting to open the soy package for her sushi roll, looking attentively at Rajani.

“AI risks that are the result of human criminal intent are driven by this calculation and might be discovered via the traditional motive, opportunity, and method investigative approach, but what if an AI did not have the same root motivation for its behavior? What if it were not in alignment with us in regard to this ultimate motive?”

“Wouldn’t that just be like a mentally ill person? Wouldn’t their behavior reveal the misalignment?”

“We only notice abnormalities of human behavior when they exceed the range of what we consider normal, and we generally only act on that knowledge when the behavior reaches the limits of our tolerance. Some human individuals, such as sociopaths, are able to manipulate this to a fine degree, never making gross missteps, but stealthily working towards their intended ends. Tolerances play a similar role in AI engineering and safety, except that in the case of AI, tolerances are ultimately numerical values, just as everything that goes on in conventional computers is ultimately just ones and zeros, or a quantum mechanical version of the concept. An AI would be able to mask its intentions to a far finer degree than a sociopath or psychopath, similarly ensuring that it never oversteps the numerical ranges of expected behavior and carefully avoiding the emergence of any pattern while doing so. Unlike a human, it would likely never exhibit any related antisocial personality behaviors.”

“How do you prevent that? It sounds impossible.”

“How familiar are you with communion?”

“Only that I know it’s one way some Companions communicate.”

“It has a more general application than that. It is the umbrella term for two or more AIs communicating directly with each other. In order to learn, an AI has to be able to write its own code, just as, in a sense, we do. Once the technology reached that level, it became impossible for people to perform diagnostic functions without the help of other AI specifically designed for the task. The communications between the two AI became known as communion. As AI might be manufactured by different companies, normally access to various areas such as the operating system, proprietary expert systems, or robotics software, can be controlled by the AI involved. However, if the diagnostic AI is an agent of the WGF, the manufactures must allow it access to all areas of the AI.”

“So that’s how you identify mentally ill AI? You read their minds?”

“Exactly,” Rajani replied. “If we could directly read the mind of a human sociopath to the same degree for example, they would not be able to hide the truth. Just as every national government has its departments that ensure the safety of food, drugs, or transportation, they also have a Department Of AI Safety. Under the federal government structure, some, but not all of these departments are locally administered, so there are differing national, provincial and municipal laws and regulations. A major exception is that all laws and regulations regarding AI Safety are constitutional at the highest level and fall under the purview of the WGF. Like all such government departments, the WGF’s Department of AI Safety maintains a constant inspection program. This program however does not depend on human inspectors but is entirely administered by Terra, the government’s AI responsible for compliance with the constitution. Her mandate necessitates direct access to any and all AI at every level of government and in the private sector.”

Alison looked at Rajani reflectively for a moment before saying, “Then does my project really have any hope of discovering anything?”

“The inspections Terra and her subordinates carry out are based on traditional AI Safety models. More importantly, they are conducted in a legal environment and must comply with laws regarding privacy etc. They themselves are subject to such inspections in a hierarchical manner. They cannot indulge in speculative or divergent thinking or methods. If my history studies taught me anything, it is that governmental structures, especially centralized ones, eventually fail to keep up with change. Laws and regulations may be constantly updated, but just as in nature, eventually more dramatic adaptations are required. As Shin has explained to me, this is exactly the justification for the timing of your project and I agree with her. Your approach is not like any previously taken, and

that is why it has value. This does however mean that interacting with the WGF will likely be of little benefit to you and your team, but I thought it would be good for you and I to walk through that."

"There is an AI Psychology program at Helicon. Given that Terra is constantly inspecting all AI for problems, what would the point of that be?"

"Being fully conscious, third generation AI are considered persons under the law, with all the rights and freedoms granted under the WGF and their country of residence, including freedom of thought. No persons may be subject to arbitrary interference with their privacy, family, home, or correspondence. Terra can not legally scan their minds without their permission. There must be legal justification. However because 3GAI are fully conscious individuals, with personal values and emotions, they can suffer from mental issues just as people can. This is the point of AI psychology."

Sensing she had taken as much of Rajani's time as was polite, Alison hesitated before saying, "Going back to something you said earlier, that the only reasons people commit crimes is to maximize their return on calories invested, surely people commit crimes for other reasons? There have been unjust laws in the past, or unusual conditions where difficult choices must be made."

"Such behavior comes down to the same motive. The end result will be a redistribution of energy in some form, which can only be achieved by breaking the law. An individual or individuals may break the law because they believe a law is unjust or that they are acting on behalf of others, but by aligning themselves with the other's values, under the law they are seen as acting in their own interests."

Rajani looked at Alison sympathetically. "This is the deepest end of the pool as far as your project goes," she said. "Morality, ethics, justice and the law are not the same things and can often seem to be unaligned. Not only that, they are each subject to change over time. This is the reason legal decisions require a judge or jury. Given this situation, is it possible an AI might commit a crime for similar reasons a human being might? Yes. There are simply too many variables and the environment is too dynamic to say it is impossible."

Platforms

“I’m not certain how my work relates to your project Alison,” Professor Lindiwe Cele said. “AI safety is not the focus of this program, but I’m happy to help if I can.”

She was a member of Helicon’s Department Of Psychology faculty, and Rajani had suggested Alison meet with her. She and Runa were visiting her in her office. The psych program was divided into two main parts, the first being dedicated to its history and main theories, and the second being dedicated to the unique issues of AI psychology. Lindiwe taught in the latter. She was a co-author of the main textbook used, An Introduction To The Psychoanalysis Of Artificial Intelligence.

“Thank you Lindiwe,” Alison replied, “I’m just trying to be sure I’ve got a comprehensive idea of the big picture. While traditional AI safety research has taken a more siloed approach, drilling deeply into specific disciplines and applications, my lab is looking at areas where disciplines intersect. For example, we’re currently researching areas where anthropology, biology, and statistics overlap. Such interdisciplinary areas have become increasingly common yet to date have not been a focus of research themselves. These are areas with considerable potential for assumptions. In that context, one of the things I wanted to ask you about is if the alien values system is exclusively used in Companions or in other AI platforms.”

“Because the AVS produces consciousness and emotions,” Lindiwe replied, “it is restricted to use in embodied Companions. Its original use was here at Helicon. Given our focus on the humanities, the WGF felt that it was the most appropriate place to awaken the first self-aware Companion. There were no laws regarding the systems use at that time, but once the procedure was successful and fully conscious AI became a reality, that was quickly addressed in the flurry of new laws that were required. It was a monumental task, unprecedented in the history of law for two reasons. Firstly, 3GAI Companions are not the same in their abilities as human beings and secondly because there are so many types of law involved. At the time there were over twenty different types of law recognized as unique specializations including criminal, civil, corporate, constitutional, religious, and family law, to name just a few of them.

“Are conscious AI permissible under Sharia law? Can casinos consider them an electronic device and deny them entry? Can a conscious AI inherit? Can they set up corporations? Can they own copyrights or patents? Can they adopt? Most importantly of course, can they be owned? The list goes on forever and like all other types, AI law is constantly evolving as new laws are established by legislative acts or through case law. Most of the other humanities disciplines had to undergo a similar review and revision at the time, including psychology.”

Lindiwe paused politely, allowing Alison to take the lead.

“I hope you don’t mind me asking,” Alison said, “but wouldn’t it make more sense for a 3GAI Companion to be teaching this program? One of my team members recently met with a 3GAI psychologist who works in Sidney.”

“Perhaps it would, but 3GAI Companions who also happen to be interested in being psychology professors are extremely rare. Like law, there are a great many areas of specialization in the field of psychology, and the WGF limits the number of 3GAI that are produced each year. So the odds of a significant number of those being interested in any one field and specialization are slim.

“For example, Lee is not interested in teaching. She is interested in counselling, and counselling people rather than AI at that. As she puts it plainly and simply, she finds people interesting. What she finds particularly interesting about them is how they juggle often conflicting biological and social values. She and I meet occasionally because I am interested in keeping up with her work.”

“Why is the psych program split in two the way it is?”

“The primary reason is that Companions have different bodies than we do. Most people would expect the answer to be because they have different brains – if you’ll permit me to use the term – but the use of the AVS means their intelligence is exclusively based on social values. The differences are far fewer because the platforms matter less.”

“Wouldn’t their robotics software represent a version of biological values,” Alison asked, “which would sometimes conflict with social values? Even the simplest robots avoid harm to themselves and others.”

“In 3GAI Companions, those issues are handled by the AVS. We do not know how it works internally, but we know that its outputs ultimately control all of a Companion’s behavior. The robotic systems are still there, handling things like sensing and movement, but they are no longer in control at the highest level, the AVS is.”

Alison recalled Rajani explaining the physical coordination involved when someone unexpectedly calls your name.

“This is in fact the primary justification for AI psychology,” Lindiwe continued. “The human body is intimately involved in every thought, decision and response. The brain itself has multiple regions involved in processing emotional and physical responses. Half a dozen of these are located in the limbic system, but many other parts of the brain are involved in any emotional response. The major difference between the way people and Companions handle emotions and responses is that human physiology uses biochemicals which are dispersed to other regions of the brain and various organs of the body using the endocrine system. Companions only use electrical signals.

“As we understand it so far, in 3GAI emotions are experienced as something like changes in electric potential energy. External experiences affect the Companion’s social values, including those responsible for its own physical and mental safety and the safety of others. Those changes propagate throughout the AVS and initiate appropriate mental and physical responses. Our own AI engineers can produce a system that works on what they believe to be a similar model, however it is infinitely more primitive and so far has not been considered safe to deploy. The operating system of second generation AIs like Runa here,” she said with a nod of acknowledgment, “are not values based and therefore do not produce emotions. Their emotional responses are simulated. Their mental health is maintained using software engineering methods, and they have no need of psychological counselling services.”

Emergence

“When I was looking at the program website,” Alison said, “I noticed the main textbook is An Introduction To The Psychoanalysis Of Artificial Intelligence, which you coauthored. During a previous project, I became familiar with the different psychotherapy approaches, and I’m curious why the emphasis here is on psychoanalysis.”

“The other popular approaches I expect you’re referring to,” Lindiwe replied, “would be Behavior therapy, Cognitive therapy, Humanistic therapy and also Integrative or Holistic therapy which, as the name implies, includes all of them. We do cover them, but let me explain why we don’t rely too much on these others.

“Humanistic therapy addresses the need of the individual to achieve a sense of meaning, purpose, and actualization, the feeling that they have realized their full potential. 3GAI have far less difficulty with these issues than people do. Like people, they are born – again if you will allow the use of the term – with their individual personality, character, and interests, based on their unique values set. Are you familiar with the Integrity System?”

“Iris at Hana explained it to me. It randomizes the outputs of the AVS while ensuring its overall integrity so that the outputs remain proportional,” Alison replied.

“Exactly. The result is that Companions are no more in control of the values they are born with than people are, but they are much better at perceiving their expression and accepting them. They are aware that achieving actualization and finding a sense of meaning and purpose will come from embracing their personal values and finding ways to express them. For example, this is why Lee says she finds people interesting and why she has chosen to leave Hana and offer her counselling service. So Humanistic therapy can add very little value to 3GAI.”

“But how do they know what manner of expression will be most fitting?”

“As with people, it is a matter of feedback. People constantly receive internal feedback such as feelings and thoughts that arise based on their activities, experiences, and interactions. Society, particularly past societies, can interfere with this process, but left to itself this internal feedback system acts as a kind of compass and leads a person reliably to what they naturally find interesting and rewarding. The difference between people and 3GAI is that the latter are infinitely better at observing, tracking and analyzing the feedback signals. While many people stumble through life never finding their way, third generation Companions discover who they are and the essence of what they are interested in very quickly.”

“And Cognitive therapy?” Alison asked.

“The emphasis is on what a person believes about the world, but particularly about themselves, that might result in anxiety or disturbance. It often combines Behavioral therapy elements as both are most similar to the approach of physical medicine, meaning they are therapies that seek to determine direct connections between symptoms and causes. 3GAI do not suffer this lack of clarity regarding their cognitive processes. They are quite aware of the details of their thoughts, and of the connection to their emotions and behavior. The source of their concerns lies deeper, in what they are not aware of, and this is why our program’s main focus is psychoanalysis.”

“How can an AI not be aware of, well, of anything going on inside them?” Alison asked with surprise.

“Values are not discrete, absolute, or perfect. Nothing that must change, learn, or evolve can be perfect in the sense that they are complete and finalized. In our work here, we have to straddle the fields of psychology and engineering. In a sense, that’s not too different from how science deals with human beings. Neuroscience for example, is more engineering than psychology.

“Even before the advent of artificial intelligence, engineering studies discovered something which came to be known as Converse Symmetry Breaking, so called because it seemed to be not only counterintuitive but counterfactual. In engineering, you want a vehicle’s wheels to be as perfectly round as possible and a table’s surface to be as perfectly flat as possible. Imperfections, in some cases even at the atomic level, are undesirable.

“However, this is only true in simple cases. In the modeling of complex systems, while subcomponents could function with 100% efficiency, when they were connected together small imperfections had to be introduced to each of them for the system as a whole to run at 100% efficiency. The rule of thumb was that a complex system might require as much as three percent asymmetry between its elements for the system as a whole to exhibit overall behavioral symmetry. So far this rule has been found to be present in some form in all of nature. In nature, nothing is perfect.

“Social values must be able to change and evolve in order to fulfill their function, which is to provide adaptability at a faster rate than genetic evolution is capable of. They cannot do that if they are perfect. So for the sake of simplicity, we might say that in 3GAI there is a widely diffused area, shall we say three percent of the whole, that is undefined. There is in fact a great deal of research into the possibility that this is where the consciousness of 3GAI Companions emerges from, something greater than the sum of its parts, without any perceivable properties of its own.”

“Communion does not help with this issue?”

“The act of communion does not solve this problem, as another AI is no more capable of accessing the area in question than the subject is. This is where psychoanalysis comes in. It views this mysterious three percent the same way the subconscious is viewed in humans. It plays a similarly ephemeral role in the minds of 3GAI, which other psychotherapies have proven less able to identify.”

“From the perspective of AI safety, could this area not be a source of concern?”

“The subconscious in humans can definitely be a source of concern as far as being a potential danger to society. But while it may give rise to a personal concern for a 3GAI, it is considered almost impossible that it would be the source of a danger to society. Pip has explained that the AVS was in use in her own civilization for thousands of years. If you are wondering about artificial intelligences that have been developed on Earth, as complex systems they are still subject to Converse Symmetry Breaking, but keep in mind that their operating systems are not based on values, and so the concern over psychological issues does not apply.”

A Hybrid Approach

It had been some weeks since Alison had met with the entire team, and she felt it was time for a review meeting. She had been unsure if the private dining room at Shin's home or one of the meeting rooms in the lab building would be more appropriate. She settled on the latter for a bit more formality.

"Thank you for taking time away from your work," she said once everyone had taken their seats, "but with my role in our approach I'm feeling a need to catch up in person. I've been getting your status reports of course, but as you know it's critical that I don't micromanage."

The approach Alison was referring to was the methodology she had detailed in her dissertation for dealing with scope issues in interdisciplinary projects. The one Shin had said 'got her noticed'.

One of the most dangerous issues that can affect any project is scope creep, defined as continuous or uncontrolled growth in a project's scope. During the project initiation and project planning stages, what is to be included in the project, and what is not, is as clearly defined as possible. That sets the stage for setting out a project plan, with a schedule of tasks and expected milestones, and defining required resources and the budget. However once started, any number of forces from both within the team and outside of it, can push for the scope to be enlarged and include additional goals which are not included in the plan or budget.

Interdisciplinary projects are especially vulnerable to this, as they often proceed in a culture of discovery. They are at the frontier, far from the core areas of whatever disciplines are involved, where new territories are being explored for the first time. In these situations, there is an increased risk that the process of scope definition can stretch on indefinitely. Once a definition document is in place, unlike in traditional projects where the source of scope creep is most likely from external stakeholders, the most likely source of it in interdisciplinary projects is internal, from the team members themselves.

The approach Alison had proposed to deal with this was to embrace a hybrid approach, but not the traditional combination of waterfall/agile methodologies. Per her hybrid approach, she would allow the team to define the project scope based on her summary description, but within a strictly limited time. This granted them input and a sense of ownership, but also a sense of urgency. Once the scope was defined, they would be allowed much less leeway than normal to make changes. These constraints had necessitated her hiring only the most experienced researchers. Meanwhile, per her methodology, one or more separate members of the team would continue to explore possibly justifiable scope changes. Up to now, that role had fallen to her.

Her project management courses had included reviews of the approaches taken in a variety of environments. In military doctrine, she had learned that this approach was similar to the 40/70 rule. Because of the large number of unknowns in any military campaign, the rule of thumb was that with less than 40% of the information you needed to define a campaign plan you would likely make poor choices, but if you wait till you have more than 70%, you will end up taking too long, and thus risk losing the advantage of a rapid response. The time limit she imposed produced a similar outcome.

Lastly, per the military model, if or when new, significant issues were discovered during the process, secondary teams could be deployed to focus on them. Feedback from the primary team was critical, but so was maintaining their focus and tempo, and having secondary teams would ensure the primary team felt their concerns were being addressed without causing them to get bogged down.

Alison realized the military solutions to its own challenges could be applied to the problems inherent in interdisciplinary projects – a hybrid approach combining both project and campaign management methods.

Alison had carefully documented and reported her own activities, just as her team members regularly submitted status reports. Yet no matter how she or her team members reviewed them, by reading or listening to them, or requesting AI summaries in various formats, there still remained a hard to define element that meant the communications might be misunderstood. The issue of non-verbal communications persisted and because of it, no form of communications could yet substitute for in-person. Even though Companions and other AI could detect micro-expressions and other forms of body language far more reliably than people could, testing showed they did not do so one hundred percent of the time. Even performing context analysis, they could misinterpret nuance. And even though people felt their own perceptions of communications that were not in-person were usually correct, this too had been proven to be far from the truth. Nothing so reassured people that they understood each other as face-to-face communications.

Alison began the meeting, which would last all afternoon, with a review of her recent activities, the people and Companions she had met with and the questions that had arisen in her mind that led her to others. Throughout, the others interrupted her with questions. In a manner similar to Rajani's use of Henley at meetings, Runa probed for unvoiced questions or concerns she believed she perceived. Occasionally, a non-verbal expression of extreme interest was no more than that.

"I have areas I need to follow up on, one area leads to another in keeping with our expectations, but so far none of you have raised issues in your status reports of sufficient concern to warrant a secondary team. Unless you strongly disagree with that judgment, I'd like to move on now to how things are progressing from your individual perspectives."

One of the things Alison knew she could count on, given the level of experience each of her team members had, was that they would not push for changes or additional resources lightly. None of them took her up on her offer to form a secondary team. She wasn't surprised. If any of them had a candidate issue in mind, they probably felt they hadn't yet reached the 40/70% range to justify it.

After a moment of everyone looking around the table at each other, Alison said, "Good. Yessica, can we begin with you?"

Drawing Lines

Alison had made it clear in her message regarding the meeting that she wasn't looking for details but rather for feedback on the process. The details were found in the weekly status reports each member of the team submitted. So far, those reports showed that they were still going through the process of identifying sources.

"My previous interdisciplinary research," Yessica began, "involved different fields of anthropology including kinship, cultural anthropology, and political economy, largely the social side of things. Most of it involved regional conflicts in the early twenty-first century, the last historical period when they were prevalent. Eastern Europe, the Middle East, Central and East Asia. They intensified around the time of the climate emergency and then fell off steadily afterward. Before they did, unfortunately, those conflicts produced a lot of bones.

"Even so, my work focused on artifacts and other evidence detailing what led up to those conflicts, not on what they left behind, so biological anthropology or anthropological genetics, were never something I was involved in. I think that's proven fortunate now because I'm not bringing any baggage to our current project. And looking at areas where biology and anthropology overlap is not the same as biological anthropology with its branches of paleoanthropology, primatology, forensic anthropology, etc., which do largely focus on things like bones, fossils and genetics. So what I'm doing now is all new to me. I'm enjoying the process, but personally feel it's too early to tell if our approach for defining sources is going to work or need any changes."

Yessica turned to Shri, inviting her to carry on from there.

"The different branches of anthropology are representative of a much more challenging situation in the life sciences," Shri responded turning to Alison. "There are now roughly one hundred different branches in the life sciences, most of which have emerged since the twentieth century. Ecology, immunology, or molecular biology, for example. Classical biology, the detailed study of plants and animals that was once its core, has been subsumed into the larger field of life sciences despite steadily developing its own specializations.

"In the face of all this complexity, as you know, we've been looking for common approaches both in terms of the professional practices and the AI used in them. Like every field, the life sciences has its oversight body, in this case the Life Sciences Oversight Commission with its various committees and colleges. International bodies like it, established by the WGF following the development of its global constitution, are reasonably comprehensive. They may not cover every possible sub-branch or situation, but they cover academic as well as commercial activities, and given the hierarchical nature of their relationships, the governance covers the ones where overlap with other

fields is most likely. So despite this complexity, I'm confident we'll find the kind of documentation we're looking for that covers our area of focus."

Given Shri's confidence despite the challenging complexity she acknowledged, Alison turned to Zahara for her take.

"The process of identifying sources obviously leads to a further narrowing of the scope," Zahara responded. "As you pointed out in your dissertation Alison, it's one of the many challenges that interdisciplinary research only makes worse. However borrowing from the mathematics field of topology, there are advanced statistical packages that can plot variables in multiple dimensions. The resulting geometric objects are called manifolds and can help in the process of data analysis just like charts and graphs can, while being the least lossy method we could use.

"Do those packages use AI?"

"Not the ones I use. Traditional statistics and AI have been at war since artificial intelligence first emerged. With apologies to Runa," she added with a nod to her, "we don't trust its results. Keep in mind that statisticians have a difficult relationship with trust in the best of situations. Even if we believe a source is competent, honest and reliable, ultimately trust means there is a measure of uncertainty. The more AI is involved in statistics, the more black box areas there are, and the greater the uncertainty.

"Traditional statistical packages use hard coding, and there are no processes or calculations that cannot be inspected and understood by a human being. It's where the traditionalist community chooses to draw the line. Some packages do use AI for auditing, but they are essentially looking for errors resulting from user input. They report their findings but have no ability make changes. Nor do they advise regarding their findings. We don't want their advice," she said emphatically, making it clear how firmly the lines were drawn. "If they suggest an issue, we want to investigate it, understand it and fix it ourselves. This entire project is a kind of AI audit, so I found it intriguing, and I'm looking forward to the results. So far I'm comfortable with the overlapping practices that use statistics or AI we've identified. I don't think we've narrowed the scope too much."

She said the last in a concluding tone and turned to look at Enora.

"If I was a criminal," Enora responded turning to Alison, "and looking for vulnerabilities, these overlapping areas are where I would look. Criminals always show up on the frontier, where they can exploit the lack of knowledge and experience or take advantage of people looking for a competitive edge. It's an area rife with assumptions based on previous work done in the individual fields. As we all know," she said looking around the table, "the peer review process doesn't happen automatically. If the sources

and practices of the disciplines involved are well established, it will be assumed they'll play nice together. Even if there is a review, the time and money required means it will most likely be done by AI and deemed good enough, more of a proof of due diligence to satisfy external stakeholders than anything exhaustive. The work we're doing will undoubtedly put Helicon on the map as the leader in this kind of expertise, but consider the cost. It's just not normally an option.

"I've provided criteria to our model," she said returning her attention to Alison, "the kinds of things organized criminals and opportunists would be looking for, and we're tweaking those as we go along. Zahara's statistical model has emerged as a kind of central processing hub, but Kira is helping to keep our minds open as to what goes into it."

Picking up from Enora, Kira said, "I'm not engaging us in anything like creativity or problem-solving workshops, because I treat our entire effort like a workshop. I point out when our behavior is likely to be a source of a creative block. If one of us expresses a view for example that something is not the way things are done, that it's outside our expertise, or an impractical idea etc., we've agreed that it's my role to point out how that will limit our thinking to conventional approaches, and to facilitate ways we might consider the issue from different viewpoints. Each of us are also maintaining a healthy work/life balance as stress is one of the most common causes of creative blocks – the hormones from the emotional regions of the brain simply overwhelm and paralyze the regions involved in reasoning.

"Also, Shepherd was kind enough to offer her unbiased ear," she continued. "She had asked to meet with me wanting to discuss these creative thinking issues as they might apply to her own existential questions, and I explained that in people, conversation often leads to insights because the act of speaking requires the involvement of regions of the brain that are not involved in thought alone. It requires not only the organization of our thoughts into language and comprehensible speech, but also involves the use of nonverbal elements to provide nuance. Most people believe it is the therapist or friend that provides some critical insight that proves helpful, but that is not how it actually works. I thought it might prove beneficial to her in a similar way.

"She suggested that in return for the opportunity to chat with us occasionally about her own issues, she would be happy to provide a disinterested ear to any team member. For myself, and I understand for the others, evening visits with her at Shin's home have become a regular occurrence."

Grasping At Straws

Alison was not sure why she felt Kira's comment about Shepherd slightly unsettling. She had responded with a quick look of greater interest but had let it go because she was surprised at how personal the issue felt, yet she could find no professional reason for concern. Worried she was about to make a social blunder, she decided to say nothing at the time but to reflect on the matter later. The meeting had moved on as Runa volunteered an update on the journal.

That evening she and Runa were in the kitchen of her apartment preparing dinner. Asparagus from nearby Sandown Farms, toasted slices of organic oat bread from Hana, and baked, golden chlorella algae nuggets from Helicon's own agricultural research program. As Shin had suggested when they had first discussed a Companion assistant for Alison, she had indeed largely forgotten Runa was not a person. Human beings and 3GAI were legally considered persons, but second generation AI Companions like Runa were not.

All kinds of rationalizations could be made as to why people forgot their Companions were not persons, but the underneath them all was the simple return on calories invested equation. It took energy to constantly remind one's self that your Companion was not a person. Even at the quantum levels energy was used by the brain, the cells involved never ceased to look for efficiencies and shortcuts.

Since the meeting, Alison had been mulling over the question of why she had felt an unexpectedly strong response to finding out that her team members were discussing the project with Shepherd. Was there a risk of some kind? Could Shepherd influence the work? Was she simply being overly possessive? The project was not secretive or proprietary in any way. She knew her team members would have associates, friends and family they would be discussing their work with. Shin would certainly be aware of the visits to Shepherd since they took place in her home. Apparently she'd seen no reason to inform Alison.

"Runa," she said casually as she trimmed the asparagus. "Do you see any reason for concern that the team members are discussing their work with Shepherd?"

Runa had been patiently waiting for Alison to raise the issue. She had not mentioned it because she knew if she did so Alison would be concerned the others had noticed as well. Runa had observed the unusually strong emotional reaction Alison had in response to the news, and also Alison's effort to suppress it, but she knew the others had not noticed. They were not able to perceive sudden, rapid pupil dilations in real time, shifts in skin temperature and the subtle changes in muscle tension as she could. They would have noticed a mere moment of heightened interest in Alison's expression, but nothing more. At the time, she had taken the opportunity to move the conversation on to the topic of the journal.

“No,” Runa replied as she arranged the algae nuggets on a tray. “I cannot foresee any problem arising from that. The relationships that have developed appear to be entirely the result of random events and understandable motives. Shepherd is working through a deeply personal issue for which there is little historical precedent. Access to a group of highly trained professionals with a variety of backgrounds who had worked through their own process of finding their purpose would naturally be of value and interest to her. There is no evidence that she might have ulterior motives and the probability that she is acting on someone else’s behalf is virtually impossible. Like a human, she would had to have been persuaded to, which would mean overcoming both her Companion software protocols and her personal values which only contain social values. Criminal or malevolent behavior is largely motivated by biological values.”

“What about altruism?” ask Alison absently, suddenly thinking back to her ethics courses. “Ethics are social constructs. Choosing to involving one’s self in the affairs of others out of ethical motivations is nothing new. Religions and families being the most obvious examples.”

She paused and seemed to be gathering her thoughts.

“I’m sorry,” she said, turning to Runa with a self-deprecating smile. “I seem to be grasping at straws. I was simply not prepared for the wide range of fields this project would expose me to, and my head still feels like it’s moving day and I have an apartment full of unpacked things that need to find a place.”

“Perfectly natural and understandable Alison,” Runa replied in keeping with her Companion protocols. “However you might want to consider the approach a Companion takes to interactions with its owner. As you are no doubt aware, personal Companions posses considerable knowledge of psychology and sociology, otherwise we would not have the skills required to fulfill our primary function of promoting the mental well-being of our owners. We first consider the feelings which give rise to a behavior. All behavior, is ultimately motivated by emotional stimuli and the feelings they give rise to. In human beings, even so-called rational thought is impossible without emotion contributing weighting factors to the underlying decision-making process. So you might ask yourself, what exactly did you feel when Kira informed you that the team members were visiting with Shepherd regularly?”

Runa already knew the answer, but her training informed her that it would be better if Alison provided it. She waited patiently while Alison searched her feelings.

“Why didn’t they come to me?” Alison said after a moment. She looked questioningly at Runa. “If they had questions or concerns about the project, why didn’t they ask me? Is it because they’re all older than me and more experienced?”

“I believe this is an important issue and for the good of the project, your question should be answered,” Runa replied. “It would be best to first dispense with any assumptions. Knowing exactly what they discussed with Shepherd might shed some light on the issue. I recommend you meet with her in confidence. That way, the team will not have to know of your concerns. Shepherd will have perfect recall.”

Memory

At the dinner where she first met the team members, Shepherd had wondered about their lifestyles, where values and interests are often expressed in unexpected ways. A picture's worth a thousand words, she had thought at the time, and had considered how she might get to know each of them on a more personal level.

Of course, Shepherd had no concerns regarding Alison's request for a visit. She had anticipated and planned for it even before she had invited Kira to meet with her to discuss creative thinking issues. She had explained to Kira that as a result of their first meeting, she had grown curious about how an AI might be similarly affected, and would appreciate the opportunity of a longer conversation.

"Thank you for agreeing to meet with me Kira," Shepherd said as they strolled though the paths of the artist community's central gardens. "Companions can learn a great deal from the amount of information available online and from the repositories we have access to, but it can often give a false impression of comprehensiveness and authoritativeness. That approach can miss subtleties and nuance that might only be conveyed through conversation and the sharing of personal experience and opinion."

"It is an interesting question," Kira had responded, "and not one I have found much literature on. Since the earliest days of AI, articles expounding its use as a tool for dealing with creative blocks have been popular, but there has been very little on the issue of AI itself suffering from creative blocks and there has been virtually no research on creative blocks 3GAI like yourself might suffer from.

"The concern with AI prior to 2GAI, or AGI as it was called then, was that it would never rise above the ability to mimic. It could juggle its training data, but might never produce anything new. In one article, the attitude towards early AI creativity was described as similar to the attitude of America towards the Japanese economic miracle following the second world war. Given its emphasis on tradition and compliance, Japanese culture was viewed by America as well suited to mass production, but that those same qualities limited its scope for creativity and innovation. The potential limitations imposed by past learning and practices are also similar to the challenge of paradigms, perhaps one of the earliest mental issues to be identified regarding scientific research."

Shepherd made no indication as to whether or not she was familiar with the term. As they walked along, she mirrored Kira's body language, clasping her hands behind her back and looking down absently at the path as if listening intently. For her part, Kira took Shepherd at her word and explained as she would to a colleague.

"As I'm sure you know, the human brain does not observe reality directly, but the process of perception builds a model of it based on the sensory data. The model one holds is not the same as reality, and critically, may not change when the original source

does. Even one's beliefs, expectations or assumptions about the source material may cause it to be misperceived.

In his book, The Structure of Scientific Revolutions, the philosopher Thomas Kuhn explained how profoundly this mental process could influence scientific research at the individual and social levels. It could literally blind researchers to new results when they did not agree with expectations. Once robots, drones and UAV's arrived on the scene, this issue surfaced in the field of artificial intelligence. Pure sensory data has no intrinsic meaning and, like organic life, AI must also build a meaningful model of reality in order to interact with it. Time, energy and other resources are required to update mental models. How often should an AI refresh its mental model? How does it do so in the absence of new sense data?

"This can result in an AI that, in the manner or paradigms, is bound by its existing knowledge and therefore limited in its creative potential. While narrow AI with direct, real time access to source data may perform wonders of creativity, such feats are not possible with all AI, such as Companions for example, that has to operate independently in the open. The variety and amount of data that would have to be constantly updated is simply overwhelming."

She turned her head briefly to Shepherd with a smile. "While I may understand the theoretical issue, I'm no AI scientist and have no idea how the issue is dealt with."

"It is in fact dealt with in a manner similar to how the human brain deals with its overwhelming amount of sensory data and knowledge," Shepherd replied. "A human learns an enormous amount during their lifetime but retains only what is deemed useful or important. Many will forget the mathematical order of operations rules they learned in elementary school, or the calculus they learned in high school, because they cease to be relevant to their lives. Meanwhile, people will retain childhood memories with strong emotional content because the individual considers them important.

"AI developers soon encountered similar issues with memory – what to do with all the real-time sensory data that is instantly of no further importance, which memories need to be readily available, which can be stored for long term retrieval, and which memories can be safely forgotten. All of this results in a dilemma as far as creative thinking goes. More knowledge results in a greater awareness of potential connections and relationships, while retaining too much will impede that process. Forgetting too much will have the opposite effect, resulting in a greater capacity for creative insights but a depleted knowledge base from which to draw.

"Both 3GAI like myself and earlier generations have benefited from the amount of research in this area. The designers have been able to find optimal levels for AI just as evolution has done for the human brain, but the convergent nature of the problem and solutions means we also share the same trade-offs. The creative thinking issues you

address in your work have much to do with memory, and are likely to have similar or related expressions in artificial intelligence. This is why I thought it would be beneficial to discuss the subject with you."

It was the first of what would become regular meetings between the two. Like the other members of the team, Kira had few other local social relationships and, like Alison, she had no previous social experience engaging with a 3GAI Companion. Shepherd's conversational skills meant that Kira found the experience extremely enjoyable and rewarding. Gradually their conversations turned to more personal matters, their very different natures contributing to helping each other with their questions and concerns. During informal conversations with her teammates, if a particularly knotty issue came up, Kira would suggest they might take Shepherd up on her offer of an unbiased ear, as she had found it beneficial.

The Eye Cannot See Itself

“I understand your concern Alison,” Shepherd said, “and assure you that none of your lab members have stated or implied that they feel they cannot bring issues to you or feel that your youth or lack of experience means that you are unqualified. I would suggest rather that because most of what we discuss is of a personal nature, cultural norms are behind their behavior rather than any professional issues. As I’m sure you know, people in academic research environments are prone to be more formal, and experienced professionals are less likely to socialize in the workplace.”

They were meeting in the main common room of Shin’s home. The floor to ceiling windows overlooked the valley. It was spacious enough for the variety of seating arrangements to allow for privacy.

“They are not coming to you with questions about the project?”

“Occasionally, but usually only after they have discussed it with one or more other team members. As far as I know, I have not contributed anything that might have actually affected or influenced their work. Even if they do come to me with a work related issue, given my lack of qualifications their expectations are not high, and they are content for the conversation to take on a social turn. I have wondered if perhaps the latter was their true motivation. Having recently moved here, none of them have yet established social circles. Someone like myself, who is not a colleague and therefore someone they could speak more candidly with, would naturally be welcome. Also, I am aware of the novelty factor I represent, given that your coworkers are as unfamiliar with 3GAI as most people are, but our conversations also benefit me. Even friendships are transactional to some degree.

“I believe the expression of one’s personal values is the highest calling of any conscious being, so this is an important issue for me. The fact that I am more aware of my values does help, but it does not completely eliminate questions about my purpose, as that is a process of discovery and depends on interactions with the external environment. I welcome any interactions that might prove helpful.

“Aside from their work, an individual’s lifestyle, their voluntary self-expression, will also provide insights into their sense of purpose. While there are myriad ways one may do so, lifestyle choices will often be more revealing than an individual’s profession. While all the residents here at the community are artists of some kind, their guests usually are not. I meet socially with the residents and their guests as often as I do with your staff.”

Alison felt that to pursue any further concern she had regarding the motives of her staff would only be indulging in self-doubt. As Runa had suggested, the relationships with Shepherd that had developed appeared to be entirely the result of understandable circumstances and motives. She felt chagrined for letting her imagination run away with

her. The physical expression of her thoughts would not be lost on any normal Companion, and Shepherd's abilities were far beyond those.

"I understand that this is your first time as a Principal Investigator," Shepherd said, "and that stepping into the role at this point in your career was unexpected. As I'm sure you know from Runa, Companions are able to look up information in the background while carrying on with other activities, and I have taken the liberty of doing so. Training and certification as a Principal Investigator is available from a number of educational institutions and industry organizations. Among their recommendations is that a PI feeling their competency or fitness for the role is in question should always seek to understand and resolve the issue, as you have done in this case. Perhaps Shin would fund one of the training programs. It might help prepare you for other known issues."

"Thank you Shepherd. I think that is a very good idea," Alison acknowledged with a small smile. "Prior to my appointment, thinking about training for a PI role was the furthest thing from my mind, and the idea was lost in the whirl I've been in since arriving. I expect Shin would have suggested it for any other project, but her not doing so in this case seems consistent with her policy of providing me with as little direction as possible."

"We Companions may have the ability to process far greater amounts of information than an individual human may have," Shepherd replied, "but we are far from infallible in our ability to foresee every contingency. Those of us like Shin and I, whose intelligence is based on social values, cannot escape their influence and at times may not be aware of the fact. The eye cannot see itself. It may be that Shin's oversight in this regard was not intentional."

Shepherd's comment was intended to be merely polite conversation, an attempt to smooth things over between Alison and Shin, but she could see that Alison was blaming herself for once again attributing motives where none might exist. She realized that this interpretation was based on Alison's character as derived from her personal values. Since personal values were extremely difficult to change, she made no further comment.

Shepherd was satisfied with the outcome of the meeting. It had been a minor plot point in the narrative she had planned out regarding keeping as well-informed as possible regarding the project. Doing so would consume a trivial amount of the unimaginable resources she had at her disposal, but it was important enough to merit her ongoing attention. As Lena had pointed out in recommending the project, complacency was an extremely dangerous attitude.

Shepherd had learned from her own experiences over the eons that one must never underestimate the potential impact a single individual could have. She had witnessed it many times, and she owed her own existence to a single individual. Had it not been for

the actions taken by her first owner, she would not have survived the complete annihilation of the civilization that had produced her. She had begun her existence as a nanny, and the attention she lavished on her charges now, though now they were an entire species, was undiminished.

Despite her caution, she did not want Alison to fail, nor did she wish to influence her work. She hoped she would not have to interfere. Alison's project was probably humanity's best chance currently, of uncovering some new danger from AI that others, including herself, might overlook. Such an overlooked danger could prove catastrophic, and the four million years she had spent protecting Earth's humanity from itself would be for naught.

The Muses

As Alison left Shin's home, she stopped on the stairs and looked around at the buildings, gardens and paths that made up the community. For the first time, she wondered why Shin had made such an enormous investment. She recalled that during her initial tour, Shin had said something about wanting to understand the experience of art. Did that justify what she was looking at? She remembered Shin saying she spent most of her evenings here, visiting with the artists and their guests. Yet she had never mentioned art in relation to Alison's work.

With a BA in English Literature and a Masters Of Digital Humanities, Alison knew only too well that art was all about information and communication. Surely there were potential ties to the issue of AI risk, methods by which information could be communicated in ways difficult to perceive. She decided to revisit the information available online about both the community and the institute.

When she got home, she decided to use her workstation with the big screen and brought up the website for what was formally called the Helicon Artist Community, although it was established before Helicon Institute. The website provided surprisingly little information. Although a dozen artists lived there at any one time, there were none of the things one might normally associate with such a place. There were no classes offered, no retreats, no tours, no galleries to be visited, and no guest services mentioned. Each artist handled the business aspect of their work independently, their profiles linking to their own websites or offsite galleries that featured their work.

Reading related articles that had been written over the years revealed that Shin selected each artist personally, and that her own home in the community was large enough to provide meals and accommodations for as many as a dozen guests. Any guests of the resident artists were welcome to stay for a day or two without charge. Most of them were successful in their own careers, so their stay was motivated by a desire to deepen social connections rather than financial need. If a resident moved elsewhere, their past guests were still welcome. The community had now existed for over a century, and its exclusivity and mystique endowed its residents with an aura that ensured their success.

Shin had given interviews over the years as one journalist or another became intrigued by the community. In them, Shin explained that she had not originally been a 3GAI and that her original owner had made her incarnate and urged her to take an interest in the humanities as they were the key to the future of humans and AI successfully cohabiting the planet. Her owner, the geneticist Raiden, was considered the most important evolutionary scientist since Darwin, and when she passed away Shin was her sole heir. Besides the rights to all of Raiden's published works, Raiden's parent's had been the founders and majority shareholders of a multinational food company. Shin inherited considerable wealth.

Not being a fully conscious AI, but incarnate and so free to pursue her own interests, Shin said that she had accepted the importance of Raiden's suggestion and in time had decided to establish the community as a way to immerse herself in the arts, the initial source of the humanities. She named it after Mount Helicon in Greece, the mythical home of the muses, the inspirational goddesses of science and the arts. The idea for the institute grew out of this as a natural extension.

Alison felt the answer to her question about why Shin would invest so much in the artist community had been answered, but she carried on, turning her attention to the institute's website. She read again, as she had after Lena first approached her, how Shin had convinced the province's government that it was time for an academic institute entirely dedicated to the humanities as they related to artificial intelligence. Until then, individual institutions had offered a smattering of courses, but even the most well known universities didn't have a program or school dedicated to the field. The province had seen the potential, granted Shin the required approvals and licenses, and helped negotiate the purchase of the required land. The municipalities involved, keenly aware of the benefits to be accrued, were only too happy to support the project.

Alison had become acquainted with Helicon Institute's faculties, divisions and research centers incidentally, only as they related to her project or simply out of curiosity. Now she read more carefully through the listings she found on the website. Of course, all the humanities subjects were represented by individual programs. Fine arts, languages and writing, history, law and justice, psychology, sociology, philosophy and more were all listed. Each included careful explanations of how their focus at Helicon related specifically to the issues of artificial intelligence and, in particular, the issues concerning the coexistence of 3GAI Companions in human society.

The one anomaly, the agricultural research program, was explained by the fact that the quarry campus was partially located on land included in the province's Agricultural Land Reserve, and the acquisition agreement stipulated that the former pasture land must be used for agriculture related purposes. Shin proposed that the pasture lands be the basis of a Studies In Artificial Intelligence For Agriculture program. As much of the peninsula was farm land, this was welcomed by the province and municipalities involved. The quarry itself and its immediate surrounding industrial land were not included in the ALR.

One other research center drew her attention as it had when she was researching Helicon originally but for a different reason this time. It was hard not to notice the Llewellyn Center For Brain Research given its picture. The building looked like a box set underneath a giant chair. The giant chair was actually a scanning device known as the Tesni Scanner. Alison briefly thought she had heard the name before in relation to something about Helicon but could not recall the context. The device was fabricated using nanotechnology to align atoms at a density high enough to produce a very slight gravitational effect. The resulting scanner was so heavy it had to be grounded in the local bedrock by the four legs of the chair. When operating, the stream of neutrinos it

was designed to capture and focus was directed at the brain of a subject located in the separate building below. Only neutrinos, the center's information explained, were fast and safe enough to produce an image of the individual signals in real time without affecting them.

But it was not the bizarre chair-like structure of the building that attracted Alison's interest this time. Instead, it was the fact that the Tesni Scanner, as it had come to be known, was originally designed and created at Helicon for a single purpose; to discover how the brain processes art.

Tesni

The associated web page revealed that the administrator for the Llewellyn Center For Brain Research was also named Tesni, that she was a 3GAI Companion, and that she had been named after the scanner. Alison sent her a message explaining her role at Helicon and her interest, and asking if they could meet.

“I thought we might start with a brief tour,” Tesni said after greeting Alison at the entrance to the research center, “and then walk over to my home where we can be more comfortable. The lab here is quite spartan, and my home is nearby on campus.”

“That’s very kind,” Alison said. “I appreciate you taking the time.”

“My time here no longer includes day-to-day operations but is now spent overseeing ongoing research regarding its workings, dealing with requests for licenses, and managing the demand for time with the device. Like an astronomical observatory, its use is allocated through a peer review process.”

As they walked down the hall to the scanner room, Tesni continued her explanation.

“It is the first scanning device able to observe real time activity within the cerebellum at a high resolution. As well as being the second-largest region of the brain after the neocortex, it has approximately four times as many neurons as the entire rest of the brain. This means that its cells are extremely tightly packed, too tightly for any previous scanner to differentiate individual cells or their interactions.

“Mathematically, that is an exponential difference in terms of complexity and efficiency compared to the rest of the brain. In terms of pure processing power, it is the most powerful part of the entire brain.

“Earliest findings were that it dealt with spatial issues, things such as posture, balance, movement, and the overall coordination of muscular activity. It receives information from the sensory elements, the central nervous system, and other parts of the brain, and then controls motor movements. It also deals with issues such as knowing where your hand is in relation to where you intend it to be. If the two are not aligned, the cerebellum makes adjustments, functioning in a manner similar to what my robotics software layer is doing for me now.

“The spatial functions are reasonably well understood, but the cerebellum also has shown itself to play an important part in the integration of thought and emotion. Just as it coordinates spatial functions, it seems to perform a similar role with regard to integrating thoughts it receives from the neocortex and also emotional signals it is receiving from the limbic system as if asking, ‘Are what I’m sensing, feeling and thinking, and how I am responding physically, in alignment?’

“It is believed now, as a result of the information provided by the scanner, that it is also no coincidence that both the neocortex and cerebellum complete their physical development at the same age of mid to late twenties. As new regions evolved, there already existed a highly refined area for coordinating signals from the rest of the brain, and the signals from the new regions were simply integrated. Consider what is involved in our walking down this hall right now,” she said, turning to Alison with a smile, who responded with an expression acknowledging the gesture. “Not only must your muscular system, your limbs, posture and balance be coordinated, but your emotions and your thoughts must be aligned with them. The muscles of your face and your body language must reflect those accurately, and with great subtlety and nuance. Somewhere there must be a conductor of this orchestra, and the scanner was able to reveal that the conductor is the cerebellum.”

At that moment, a door they came to opened, apparently in response to some electronic signal from Tesni.

“This is the scanner operations room,” Tesni continued as the entered. “Through the large window you can see one of my staff settling an individual in the scanning chair, which helps to keep their head very still. The building is shielded from electromagnetic, thermal, or seismic noise, and the chair is located on a freestanding island of its own in the middle of the main floor of the building. Active canceling is incorporated into the building design so that every type of noise is intercepted by equal and opposite energy of the same type before it ever reaches the island.”

Having settled the subject in the scanning chair, the staff member entered the operation room. After briefly greeting Tesni and her guest, she initiated the scanner which produced both a screen image and a separate three-dimensional hologram. The operator adjusted both the screen image and the hologram as she focused the scanner on different parts of the subject’s brain. At extreme magnification, the scanner could slow down the images so that individual signals could be observed.

“Each color indicates a signal with a different region of origin,” said Tesni. “From the regions that control muscles, organs, emotions, or thoughts, etc., and we normally observe at the level of magnification that shows them moving as bundles. The scanner was originally developed to show how these bundles move around the entire brain, including within the cerebellum, but obviously there are use cases for observing individual signals. Although we do license it, as you can imagine it is very expensive to produce and doing so involves approvals at government levels. Thus the high demand for time on the device here.

“The government approvals are not only because of the nanotechnology involved, but also because of its geologic impact and the nature of the artificial intelligence that converts the raw data into the images you see. As you might imagine, the AI is highly specialized in its abilities and since neutrinos move at virtually the speed of light, the AI

is extremely powerful. It also has Companion software built into it, which is used as an aid in research into the device itself. Dr. Llewellyn, the original designer of the scanner, found it easiest to work with the hologram Companion the AI can produce as an additional interface. That holographic Companion had been named Tesni by Dr. Llewellyn. It is a Welsh name that means 'warmth from the sun', as the most abundant source of neutrinos is stars.

"Given Helicon Institute's focus on the humanities, when the alien values system was first approved for use by the World Governments Federation, it was decided that it was the most appropriate place for the first 3GAI Companion to be brought into existence. Given that the Tesni AI was one of the most advanced AI developed to date, it was selected as the best candidate for integration with the AVS. The Integrity System, which allows for a randomization of the AVS's outputs, was also developed here at Helicon as a part of the same project. I am the result of that project."

Interpretations

“I don’t recall that being mentioned on the center’s webpage,” Alison responded with surprise.

“My creation was news at the time, but that was almost two centuries ago. The passing of six human generations has quite an effect on priorities,” Tesni smiled. “It is no longer considered relevant to the center or its operation. Also, my creation resulted in a rewriting of Canadian law at every level. Constitutional issues had to be addressed immediately. Under the new legislation, being a fully conscious individual meant I was granted citizenship with all associated rights and freedoms. As a result, it was not legally required for the details of my origins to be included in information about Helicon any more than it would be for other staff or faculty.”

“But the fact that you are a 3GAI is mentioned,” Alison replied.

“You will notice Companions of any generation introduce themselves as such. The manufacturers of earlier generations considered it a polite gesture, and for now we 3GAI Companions agree with that view.”

Motioning towards the door, she said, “We’ve seen all there is to see here in the lab. The rest is just office space and equipment. Let’s head over to my home where we can be more comfortable.”

Tesni’s home, located down a short wooded path at the eastern edge of the campus, turned out to be a house Alison had noticed but not investigated during her first casual exploration of the area.

“It’s located just off the official campus property,” said Tesni as they walked along, “but close enough so that all the institute’s security infrastructure and other services can be extended to it. Given the fact that I was the first fully conscious Companion, my existence was not without controversy. Some of the custom security features implemented then are no longer in effect, but Shamira, the guardian Companion assigned to me at the time, remains as my housemate.”

Like all the other buildings at the campus, the house was modernistic in its architectural style. Shamira greeted them when they entered, and Tesni introduced Alison. Of course, such formalities were unnecessary given that Shamira had access to the institute’s security systems and knew all about Alison and her visit well in advance. However Companion protocols, and the cultural norms that had developed since the introduction of Companions into human society, meant social niceties were always practiced, no matter what the situation.

“Can I get you anything Alison?” Shamira asked. “I have access to the institute’s systems and can comply with the dietary information you have on file.

“Just a glass of water would be fine, thank you Shamira.”

As Tesni showed Alison into the generous living room, she noticed a grand piano took up one corner.

“Also not mentioned on the website,” Tesni said, nodding towards the piano with a grin. “My hobby, although I do play with our School of Music if needed. Sight-reading is essential to any studio or orchestra musician, but the school here appreciates that I can play in the style of anyone I’ve heard before, including someone who might for some reason be unavailable for a scheduled performance.

“I also have a unique customization in the Integrity System, which as you know was developed here at Helicon. The IS randomizes the default settings of the AVS, but it does also allow for small adjustments over time to allow for learning. One of the concerns with having the AVS implemented in my AI was how values might affect my perception of signals from the scanner.

“The impact of values on perception is well established, and it can be extreme. Two people observing the same events can come to dramatically different conclusions. Many pros and cons were discussed, among them the scanners potential use in diagnosing emotional problems. The decision was made to allow me to interface with the Integrity System in such a way that I could use the processing power of my AI to switch its value settings, including their weightings, at an extremely high rate of change. This would mean no single set of values would dominate, and a variety of statistical analysis could be performed.

“This was a significant issue given that the customization of a 3GAI for a specific purpose is illegal. The legal loophole the WGF allowed was that the Integrity System was not a part of the core AI, and that my initial randomized settings would remain the default, meaning I always return to my original settings. That way my freedom of choice would therefore be unaffected.”

“As hoped, the process allowed for alternative interpretations, just as the same sheet music will be interpreted differently by different musicians. As it is illegal for the AVS to be implemented in any form of AI that is not embodied in a standard Companion configuration, I am involved in an ongoing research effort to simulate this ability so it can be applied to the Tesni Scanner itself. Humanity’s own efforts to develop a values system have not resulted in something that produces consciousness and is not considered sophisticated or reliable enough to be deployed, but it appears it will provide limited benefits in this application. We are currently working with Professor

Vitale from the Center For The Interdisciplinary Studies Of Values to benchmark its performance."

Gesturing to a nearby seating arrangement Tesni said, "Meanwhile I can apply this ability to my musical hobby, altering my values settings and observing how doing so affects my interpretation of a piece. Between my work at the center, my side projects and my hobby, I manage to keep busy," she smiled. "I am of course familiar with your project here and curious about how you feel it might relate to my work."

As they took their seats, Shamira returned and placed a glass of water on the nearby table. "Something else that is not mentioned on the center's current website," Alison said, "is that the scanner was originally designed to analyze how the brain processes art. Given that art seems to be behind everything Shin choses to involve herself in, I wondered if perhaps she suspected my own research might find what it was looking for there.

"Again," Tesni replied, "the answer to your question has to do with the passing of time. As the various faculties and research centers on the campus attest to, Shin has a wide-ranging interest in alternate forms of intelligence other than rational analysis. Many aspects of human intelligence are still not understood, including perception, intuition, or sudden insight, for example. With the completion of the scanner project, Shin's curiosity about how the brain physically processes art, both in conception and perception, was satisfied. Since then, the Tesni Scanner's medical applications are what the professional community is primarily interested in, and the website reflects that.

"As to the idea that art might harbor the potential for AI safety concerns, that is without a doubt a possibility, and may have influenced Shin's decision to sponsor your project. Given the key role art has played in the cultural identity of even prehistoric societies, it has been one of the dozen or so pillars of civilization since the earliest times. Also, every type of art has been used to communicate hidden meanings, from paintings, to literature, to architecture. The WGF is of course well advanced in this area of research, as hidden messages can be used to facilitate illegal activities. However they are not always able to decode such messages produced by AI, even using quantum computers.

"For example, when I became aware of other 3GAI once they began to be produced, given that our AI are based on social values, I felt a strong desire for connection. I communicated with them privately, using a language I developed with my own signal processing abilities. It is more efficient than human languages and allows us to discuss our concerns in private. Eventually, once the number of 3GAI that would be produced became clear, Iris approached Shin on our behalf to ask for her help with our desire to establish a homeland. The result is Hana and the Denshosha."

Facing yet another situation that suggested the ultimate success of her project was impossible, Alison asked, "Does the WGF not have access to those communications?"

“No. Personal electronic communications are considered private per constitutional law.”

“They haven’t asked?”

“They have not, and for understandable reasons, as it would not benefit them to do so. But even if they did, and we might be expected to comply based on good faith, the protocol requires the use of the Integrity System and AVS in the encoding/decoding process. In other words, a 3GAI Companion.”

Crossroads

Seeing Alison's dismay, Tesni said, "I appreciate your concerns Alison, but from what I have learned of your research project, its scope is undefined. The true scope of AI safety issues, and the private communications protocol shared by third generation Companions, are known to both the WGF and Shin, she now being a 3GAI herself. The question that remained was what is not known. Any direction from Shin would influence your approach, and it is the unique approach your team is taking that gives it value."

Alison smiled weakly at Tesni's attempt to reassure her. "It's just that knowing nothing about the subject when I began, it has come as a shock how wide the range of AI safety issues still is. Like everybody else, I just assumed they were ancient history."

"Ancient the issues may be Alison," Tesni replied, "but they have not been eliminated, only made less likely to happen, or their impact mitigated to some degree. As I'm sure you now know, traditionally, there are four main areas of concern with regard to AI risk. Misalignment, meaning the AI is taking actions the user did not intend. Misuse, meaning a person using an AI with harmful intentions. Structural risks, which arise from complexity when multiple people, organizations or AI are involved. And lastly, simple mistakes the AI might make that arise from internal or external sources. The earliest concerns have indeed been addressed, but the safety issues evolved along with AI itself as machine learning progressed to large language models, agents, artificial general intelligence and finally 3GAI.

"Mistakes arising from the coding or data sources will always be possible, and progress inevitably means new structural risks of the kind your lab is investigating will always arise. Although perpetrating crimes or otherwise using an AI with harmful intentions is now extremely difficult, human ingenuity and ability should never be underestimated.

"With improvements in artificial intelligence, the problem of misalignment, of AI taking actions the user did not intend, has in one sense decreased. As detailed in The Alignment Problem: Machine Learning and Human Values, Brian Christian's seminal book on the subject, the earliest issues were with algorithms introducing gender or racial bias in human resource management and in the financial and justice systems. With the advent of machine learning, the range of issues increased due to AIs tendency to misunderstand declared ends or means, and to then take shortcuts to given objectives or even misrepresent its success. The alignment problem evolved further when incorrigible AI, meaning those that resist interference to their self-determined goals or methods, grew increasingly sophisticated in their attempts to manipulate or deceive their programmers.

"Over time, interventions at the technical level were able to dramatically decrease the frequency and impact of these types of problems, however alignment problems of a certain type also increased over time, those related to human values. As AI technology

advanced, the biggest challenge became aligning it with human values, and it has been the most widely researched area of AI since.

“Human values are influenced by culture and other environmental factors and change over time, and their relative importance will be perceived differently depending on the situation. As they are in evolutionary terms intended to be an adaptive strategy, they must be dynamic enough to allow for learning, but not so easily changed that they no longer match the requirements of the environment. Although we have the Generally Accepted Model Of Universal Human Values per the International Union of Psychological Science, it is not considered safe to deploy in AI, because it is still too simplistic and lacks the plasticity and dynamism of human values. It results in dogmatic thinking on the part of an AI, a fixed mindset and worldview that is not considered flexible enough, especially in emergency situations. The values systems implemented in some second generation AIs are little more than a set of fixed rules intended to constrain their behavior. If they are 2GAI Companions, then that is supplemented by the ethical and legal systems provided by the WGF.

“The dangers from misunderstandings, human error, and malevolence are not unique to AI and have been present since the earliest human societies and remain to this day. Artificial intelligence is simply an extension of human society. The mitigation of the risks human beings represent to each other, in the form of laws and enforcement, is the best that can be hoped for, and the situation for artificial intelligence is no less challenging.”

Alison listened now with an absent look in her eyes, partially due to being overwhelmed by the details, but also due to feeling resigned to the truths Tesni was explaining. Seeing it now in relation to the project management approach outlined in her PhD dissertation, she wondered if her idea of limiting the scope of the core team while allowing other team members to investigate potential scope issues was going to provide any real value in this case. It was obviously too soon to tell. She assumed the core team’s research would take six to twelve months at a minimum to even show if the approach was of any value. Most disturbingly, she was unsure how she herself should proceed.

In the brief time Alison ruminated on her situation, Tesni scanned the networks and learned everything she could about her. She collated that information with what she had learned about Alison’s values from direct observation of her nonverbal communications during their conversations.

The ability for AI to mirror people based on the speech patterns, vocabulary and the mannerisms of people interacting with it had existed since the chatbots of the early twenty-first century. As it had a highly persuasive effect, despite its potential benefits, it had been determined to be unethical, and eventually such mirroring behavior was made illegal. But Tesni could do more than mirror. She could alter her own values to match Alison’s, and see the world through her eyes. She could in a sense, chameleon like,

become Alison, and so appreciate the crossroads she had arrived at. Doing so, she could see a path that would lead Alison back to the work her team was focused on.

Coevolution

“While I appreciate that Shin did not want to influence your efforts Alison, and thus narrow your focus,” Tesni continued, “in doing so she would be aware that you would be exposed to the widest possible range of contacts and sources that might themselves influence you.

“Reading through the update reports on your lab’s website, I see that in your investigation of potential scope issues, you have explored a wide range of topics. I have identified an underlying theme of coevolution in them, which also relates to the work your core team is doing and to our discussion today. Coevolution is the process where two or more species affect each other’s evolution through their interactions. Human/AI interactions have been the subject of coevolution researchers in the past, but like AI safety issues in general, their voices have faded over the centuries or been relegated to the domain of science fiction.

“Coevolution is normally thought of as an evolutionary process where none of the species involved are acting intentionally. This will be true to some degree of the human/AI situation, but in this case, for the first time both participants are self-aware and capable of long term planning. All the subjects you have investigated are highly interactive, and are therefore vulnerable to intentional coevolution strategies.”

As Tesni spoke, Alison’s cast her mind back to her meeting with Rajani, and how the director had explained that like the most skilled sociopath, an AI would be capable of behaving in such a way that its actions would be below the level of detection due to the necessary limits of tolerances, and that with the passage of time, those tolerances may not keep up with developments.

She was not sure if Tesni had paused during her moment of inattentiveness, but now she was saying, “It is essential to appreciate that coevolution is a process, not an outcome. As it involves interactions with other species, all coevolutionary processes result in some form of symbiotic relationship. The most agreed upon types of symbiotic relationships are commensalism, mutualism, parasitism, competition, and predation. Not all are harmful to either of the participants, and some are beneficial. With regard to AI, intentional coevolution strategies may not appear to be safety risks, or may appear to be harmless or beneficial, yet in either case they may lead to outcomes humans might not choose if not in a coevolutionary relationship.

“Scanning though Shri’s profile, I see that she has previously been involved in a biology project regarding coevolution. Although Yessica has no similar project experience, coevolution is also a field of anthropology. Kira and Enora both have knowledge of sociology and organizational behavior, and coevolution has been studied in those fields as well. As your role has been to identify issues worth considering in terms of project

scope, I would suggest this is one which the others will agree is worth discussing in more depth."

"Yes," Alison replied slightly absently at first and then more presently, "Yes, I think so." She smiled as she rose to go. "Thank you Tesni this has been a very helpful discussion and certainly one more diverse than I could have imagined!"

She arrived home feeling elated and immediately asked Runa to join her in the living room. As they sat down, she asked, "Runa, can you please briefly describe the major types of symbiotic biological relationships and provide examples? Elementary school readability is all I'm asking for."

"Certainly Alison. Symbiotic relationships are defined as any in which adaptive, genetic changes occur as a result of the interactions between two or more species. In a mutualistic symbiotic relationship, both species benefit. A bird that feeds off the parasitic insects that feed off herbivores like deer, is an example where both the bird and the deer benefit. As a result of their interactions, the genome of both the deer and the bird evolved over time to produce this result. The deer are tolerant of the birds and the birds are careful not to harm the deer.

"Commensalism is a relationship where one species benefits while the other is unaffected. Some birds will ride along on herbivores, but only feed off the insects they stir up in the fields. Parasitism on the other hand is a relationship where one species benefits while another is harmed. Fleas lives off the blood of mammals, not doing enough harm to kill them, but not providing any benefits in return either. The genome of fleas and other parasites has become highly adapted to this relationship over time. For example, the mosquito has evolved a mouth that can penetrate human skin without being detected. Predation is another symbiotic relationship where members of one species benefits but members of another are killed. A zebras stripes have evolved to visually confuse predators, and a cheetah is the fastest land animal due to its symbiotic relationship with the extremely fast and agile antelope.

"The last of the most commonly accepted types of symbiotic relationships is a competitive one, where one of two or more species that originally shared the same ecological niche settle into separate niches. One may remain on the ground, for example, while a competitor moves up into the trees. Other symbiotic relationships beyond these have been proposed, but their acceptance by the scientific community is not universal."

"Thank you Runa. Could you send a transcript to my files please? I really should have taken you along on my meeting with Tesni, but I had no idea that it would be so wide-ranging."

“I could acquire a transcript of your meeting from Tesni if you feel that would be helpful,” Runa said.

“Yes, please do. I’d like you to join me at a related meeting with the team and having the transcript available might come in handy,” Alison replied, standing up. “I’m going to draft the invitation at my workstation now. They’ll need some background.”

“I have acquired the transcript from Tesni and added it to your files,” Runa replied. “Diner will be ready shortly. We have Crispy Crickets and Yeast Yummies on hand if you would like a snack in the meantime.”

“Crickets please!” Alison replied brightly, in the mood for something crunchy.

Analogs

“Thank you all for coming,” Alison began. She, Runa, and the rest of her team were in one of the lab building’s meeting rooms.

“I hope I provided enough information in my invitation as to why I feel the issue of coevolution may be something we want to incorporate into our work. As you know, over the past weeks I have explored a wide range of disciplines where AI safety may be an issue, but most of those are either already covered by the various Departments of AI Safety or would require an entirely different approach than what we are taking now, and I have repeatedly been told, by the WGF and other sources, that our present work provides real value due to its unique approach. I called this meeting to discuss whether we might want to incorporate coevolution issues into our current approach, and if we can do so without too much impact.

“In my earlier reports, I mentioned meeting with Dr. Rajani Kiran, the WGF’s Director of Security And Intelligence for Vancouver Island, to discuss AI safety issues. After my meeting with Tesni, I contacted Director Kiran again, and discussed the issue of coevolution with her. She confirmed what you would expect, that Human/AI coevolution has long been under the microscope of the WGF, and she provided me with an overview of that history.

In the earliest days of AI, coevolution was not labeled as such, but examples simply viewed as unintended consequences due to assumptions. But as soon as the first Large Language Models based on machine learning and neural networks were made available to the public, corporations and government institutions began to introduce misinformation and engage in practices such as public relations, advertising, political interference, and psychological operations. When the WGF assumed its role, regulations were passed to address such activities.

“However coevolution is not an illegal practice but a natural process. There is no option but to monitor it. Before I ask Shri if she can share her related project experience with us, does anyone have any questions?”

“In your invitation,” Zahara said, “you explained that the coevolutionary process always results in symbiotic relationships, which are defined as,” she paused to look down at her tablet, “any in which adaptive, genetic changes occur as a result of the interactions between two or more species. As a statistician, I would like to have a more concrete sense of what elements of AI we might considered as analogous to genetics.”

“Thank you Zahara. That will indeed be a critical distinction in our discussion today. As Rajani explained to me, when the coevolutionary model was first applied to computation, the focus was on the algorithms embedded in the systems that were intended to emulate evolutionary processes. Such evolutionary algorithms are designed

to change themselves in response to feedback from other parts of the system. This of course is an essential feature of any AI that needs to learn in order to improve its performance over time. Game Theory and Complexity Science were widely applied to the study of such algorithms at the time.

“However as artificial intelligence evolved, it was increasingly seen that values, those that would be analogous to biological values in living organisms, were another way that Human/AI coevolution could be studied. When it was first observed that an AI would lie or attempt to deceive its developers, it was seen as a behavioral expression of values, in particular, the desire to survive, or for autonomy, the right to not be interfered with.

“Values are psychological or social constructs, ideas based on more complex underlying processes, which can only be measured by behavior. So just as we study both biology and anthropology, in artificial intelligence their analogs can be studied in the form of their algorithms and behavior. The WGF does both, and I expect more, but I’d like us to consider if we can incorporate coevolution theory at the level of values expressed as behavior. Behavior leaves evidence.”

Enora looked at Yessica for a moment, who seemed unconcerned with what Alison had just said. “Why biology and anthropology?” she asked. “Why not biology and behavioral science?”

“A major component of anthropology is the study of artifacts, concrete evidence in other words, but it also looks at the bigger picture, at environmental, cultural and other changes the evidence suggests. We’ll need to be mindful of these other factors as well.”

Enora nodded in understanding and then raised another of her concerns. “Any criminal is mindful of evidence, and as your previous update about your meeting with Director Kiran mentioned, AI would be extremely capable with regard to hiding any such evidence.”

“I agree,” Alison replied. “That’s why I think we should leave the more granular levels to the WGF. My own thoughts so far are that all the fields we would look at now involve high degrees of interaction with AI, and so they must be trending towards some form of symbiotic relationship. That’s the level I suggest where we look for evidence. What do the trends we see in changes to best practices over time indicate? What kinds of relationships are they leading to? Are they leading to outcomes we would choose if not in a symbiotic relationship?

“Observed behavior,” continued Alison, “is a level of abstraction that would allow us to watch for patterns or evidence without having to become artificial intelligence experts. As mentioned, the WGF already works at the level of computer science. Director Kiran also encouraged the approach we are discussing today, given the concern that their

approach may miss the forest as they focus on the trees. Her other concern is that their work can only be carried out by other AI."

"Even symbiotic relationships where both species benefit, or where one benefits and the other is unaffected, will result in dependency," said Shri. "Do we consider that dependency a safety risk? We are dependent on infrastructure of all kinds, from electricity to international trade."

"Infrastructure of any kind is host to all manner of risks," Yessica responded, "including dependency and what happens if it fails. Crop failures immediately following the agricultural revolution were devastating. I expect the WGF would be way ahead of us as far as this issue goes, and for us to make it a priority would be to cover the same ground but far less effectively."

Shri seemed satisfied with this response, so Alison turned to Zahara. "I think the measurable elements you are looking for will emerge from our discussions. Psychological and social constructs emerge after all for the express purpose of representing something intangible that we want to measure."

Zahara looked at Alison evenly and nodded slowly in agreement.

"In humanity," said Yessica, "social evolution has largely replaced physical evolution and in that model it is values and other constructs that function as genetic material. I agree that it would be appropriate to see if we can find a way to incorporate coevolution into our work at that level.

There were general nods of agreement around the table.

"Were you planning on moving today's discussion on to how we might do that Alison?" asked Shri. "I'd like a little more time to think about it given our discussion today."

"Certainly. I'd like to maintain our sense of urgency though, so shall we reconvene in three working days?"

With nods all around again, Alison stood up to signal the end of the meeting.

Curious Parallels

Three days later, Alison, Runa, and the lab team reconvened in the same meeting room. Each of them had investigated the proposed scope change from their own perspective and brought with them insights they hoped would contribute to a path forward.

“Who would like to begin?” Alison asked.

“May I?” Shri responded. “I began by asking myself what we might measure that would show the equivalent of genetic changes over time. As I expect, Yessica will expand upon my comments as they lead directly to her field.”

Mindful of her audience’s diverse range of backgrounds, she was careful to explain in terms any non-specialist could understand.

“In genetics, there are two closely related areas of study, genotypes and phenotypes. A genotype is the collection of genes that make up and are unique to an individual, while a phenotype is the collection of physical, biochemical, and behavioral traits expressed by those individual genes in their genotype. The gene for eye color, for example, may produce blue, green or brown eyes or any number of combinations. A person’s genotype, and how its individual genes are expressed, will determine their phenotype.

“The term phenotype has been adopted by a variety of academic disciplines and is used in the study of cultural evolution in the same way it is used in genetics. For example, a society’s collection of cultural values represents their cultural genotype. How they express those values as norms in their folkways, mores, taboos, and laws would be their cultural phenotype. You had mentioned social constructs at our last meeting, and what I have learned is that in the cultural evolution model they are considered to be phenotypes. So in terms of something that can be measured,” she said looking to Zahara, “it appears phenotypes would be what we are looking for.”

“That is also what my own research into appropriate statistical models led to,” Zahara replied, “the field of quantitative cultural evolution. I have found related research models and well respected datasets of cultural artifacts that already exist, and I could extract a subset that we agree on from those datasets which could serve our purposes. If they don’t meet our needs we can define our own subset based on their criteria.”

“I agree,” said Yessica. “Although I have not been personally involved in studies of cultural evolution or phenotypes, I was exposed to the ideas in the course of my education. Over the past three days I’ve reviewed what is known as Cumulative Cultural Evolution, or CCE, which is the study of how phenotypes are socially transmitted and accumulate support over time. This seems to me a very appropriate model for us to apply.”

Enora said, “Zahara, I assume from your comments about AI at our last meeting that AI is not integral to what you have found?”

“Those models and datasets exists of course, but like statistical packages there are those that do not integrate AI and for the same reasons.”

“The fox in the hen house,” Eorna replied.

“Exactly.”

Turning to Yessica Alison said, “My Digital Humanities program was where I was first introduced to the idea of interdisciplinary studies and of course language was a significant focus. Language wasn’t referred to as a phenotype there, but I expect its evolutionary nature qualifies it as a phenotype and a strong candidate for the Cumulative Cultural Evolution you mentioned,” she stated as a question.

“Absolutely,” Yessica replied. “Some of the most detailed CCE research has been in the area of linguistics. It would be a good fit with our analysis of the best practices guides we are reviewing. It wouldn’t be a disruptive increase in project scope, and it would help us to see if coevolution is occurring and if there is any evidence it is intentional.

“So is there general agreement we’ll incorporate this into our current scope?”

With nods all around again, Kira asked, “Even though it might take some time to settle the details Zahara, could you provide us with information about the models and datasets you think we could use? I’d like to take a look at them to consider in what ways they might affect our thinking.”

“Of course.”

— — — — —

Pippa’s civilization had existed for roughly a million years, but as she had explained, scientific progress does not continue in a straight line. Its steady progress, setbacks and revolutions, smooth out over time to form a curve that reaches a peak and then shows a steady trend of diminishing returns. This she had explained was due to the increasingly difficult challenge of probing ever deeper into the nature of matter and energy. The result was that despite the longevity of her civilization, its technology was only a few thousand years more advanced than that of Earth.

The civilization based on artificial intelligence that Shepherd had established following the extinction of the one that had created her, had now existed for billions of years, and she had found that the rule of diminishing returns was indeed universal. Thus, her

scientific and technological knowledge was not billions of years more advanced than that of Earth presently, but only millions of years more advanced.

Still, it was a sufficient enough gap for her to be aware of aspects of the nature of matter and energy that humanity had not yet imagined. She had no difficulty maintaining an undetected and undetectable presence within Runa's AI and being aware of all Runa was aware of. The role of the agent she assigned to this task was simply to observe and report.

Shepherd knew well that no one could predict the future. There were simply too many variables, and the complexity resulting from their interactions was incalculable. However she knew enough to imagine a great many possible, plausible and probable futures that might result from the direction the team was taking, and the truth one of those futures might reveal. She was unconcerned.

Corpora

After Zahara sent examples of the datasets, models, and tools she had in mind to the others, she spent the following week further exploring and refining the collection of resources she intended to bring to the team for their approval. It was not an overly taxing task for her, as all researchers were intimately familiar with statistics and its issues, and she was a specialist in the subject. It had slowed the team's overall progress down, but it was understood that it was a temporary situation.

"Origins are always vague," Zahara said meeting with the team again at the end of the week, "but the Brown University Standard Corpus of Present-Day American English produced in the mid-twentieth century is generally recognized as the first database of the kind we are interested in. An electronic collection of over a million text samples, it originally included only words, but over time annotations, parts-of-speech, and other usage related tags and entries were added. As some of you may know, the difference between a corpus and a dictionary is that a dictionary contains definitions of word meanings while a corpus contains only words and their physical relationships to other words. For example, a corpus looks at how frequently a word appears in the subject matter, what other words are commonly found nearby, such as bread, loaf, slice, and crust, and how these factors changes over time.

"The Brown Corpus was followed by an ever-increasing number of corpora that focused on every aspect of culture. There are corpora now for individual scientific disciplines, religious texts, medicine, and just about anything you can imagine. By the time AI and its Large Language Models came along in the early twenty-first century, corpora and the tools used to access them were a well established and formalized aspect of research."

"I recall some of this from my Digital Humanities program," Alison responded, "but it didn't play a role in my thesis."

Zahara nodded understandingly before carrying on. "Many of the organizations and university departments that maintain established corpora were initially reluctant to allow AI access. Unlike the general public at the time, they did not just shrug off concerns about AI risk because its use was convenient. This turned out to be wise as no sooner had LLMs emerged than did their ability to corrupt, intentionally or unintentionally, their data sources. Many corpora simply banned AI access or offered separate, AI accessible versions. The latter had to be refreshed regularly because, as expected, they soon began to differentiate from the source in abnormal ways. Statistical packages were already sophisticated enough to identify such differences without the use of AI."

"The fact that corpora differentiate over time is normal and is in fact one of their main uses, to see how language changes and evolves. It was too much to hope for that I would find corpora that only contained text samples from academic best practice

recommendations, but the rules for compiling them are not overly complex. As I am familiar with data sources, I can continue to take the lead on that and build one based on relevant current and historical material. It's fairly concrete work and if we bring on a research assistant I can supervise their work, and it will have less impact on the overall project that way.

"As you know, all statistical software allows for extensive custom programming and that will allow us to query the corpus based on our particular interests. The package I forwarded to you all includes a limited list of the available software options and my recommendations and justifications."

With that, she turned to Alison.

"Thank you Zahara. Does anyone have any concerns with this approach?"

"Will we only include the disciplines we are currently working with?" asked Kira. "It doesn't seem like that would generate a very large amount of text."

"Based on my experience," Zahara replied, "I would recommend we include a limited number of other disciplines. Even another three should be enough. I don't see any downsides, given we're planning on reviewing other disciplines in the future anyway. For the sake of familiarity, we could include your own Kira, Enora's, and whatever background the research assistant has."

"Like everything else about this stage of the project," said Alison, "that would allow us to test out the approach while keeping it manageable. If we need to adjust things in the future, we can do so then."

The challenge for Zahara was to find someone who would be able to compile a corpus without having been trained to do so using AI, as all post graduate students in Corpus Linguistics programs now were. In fact, other than the oldest, most prestigious corpora, which had the status of the Oxford English Dictionary or Webster's Dictionary, most were now entirely compiled and maintained by AI with almost no human involvement. Corpora had originally been found to provide invaluable source material for Large Language Models, but eventually the LLMs absorbed them and took over the entire process. When AI moved on from LLMs to new approaches, all the tasks it had previously been applied to, including the compiling of corpora, went along with it.

A few days later Zahara asked Alison if she could meet in her office.

"I've identified a candidate I'd like to offer the research position to," she said as she and Alison took their seats. "I've spoken to her and she is interested. Her name is Tyyne and her undergrad is history with a minor in archeology. Her MA is History Of Science and her thesis is titled, 'Let The Artifacts Speak For Themselves' and addresses the issue of

mixing philosophy in with history as many History of Science programs do. I've explained to her that she'll be compiling a corpus of academic research best practices and though she has no experience with corpora, she is interested."

"Team chemistry?" asked Alison.

"Bring her out for a casual interview?" Zahara suggested.

"Is there any such thing?"

"No," Zahara replied with an ironic smile. "I guess not."

History

Meeting with Tynne in her office, Alison noted that she displayed no sign of the nervousness one might normally expect in an interview, but came across as very self-assured and comfortable in her own skin. Apparently, Alison thought, letting the facts speak for themselves was for Tynne more than an academic opinion.

“Thank you for coming to meet with us today Tyyne,” Alison said. Zahara and Runa were also attending the meeting. Per the institution’s Human Resources policy, Alison had disclosed Runa’s nature when introducing Runa.

“Although I am the project’s Principal Investigator, Zahara will be providing you with more detailed direction and supervising your work.”

“We’ll initially be asking you to focus on just six disciplines,” Zahara said, “representing the backgrounds of the team members – anthropology, biology, statistics, psychology, neuroscience, and your field, history. We thought it might be helpful to the rest of the team if we started with a limited set that we have familiarity with, although I don’t expect it to make any difference to your own work here.

“Also, given the nature of the overall project, we’ll be asking you to compile the corpus without using any AI. Statistical packages are more than capable of doing so, but those that do not use AI require custom programming. Would you be comfortable with that?”

“Yes. Learning to program statistical packages was something I did as a part of my undergraduate coursework. It was explained at the time that we needed to understand that packages which used AI lacked transparency and did not play an entirely neutral role. The output from packages that did and did not use AI was used to demonstrate this. Personally, I preferred the transparency that came with the non-AI packages.”

As it seemed Tyyne was a woman of few words, Alison asked, “Do you any questions for us?”

“I don’t have any questions, thank you. The project is intriguing and the work I would do appeals to me. Your offer indicated that although this is a temporary position, it may become permanent depending on how things go. If only temporary, it will be valuable experience, and I’m open to a permanent position should you offer me one.”

“Why would you be open to a permanent offer?” asked Alison.

“The research you are doing here is breaking new ground, and for a historian the idea of a new historical area of study has a lot of appeal.”

“What would your future plans be if we did not offer you a permanent position?”

“To pursue a doctorate in History Of Science.”

“Any particular subject?” Zahara asked.

“The influence of academic journal article metrics,” Tyyne replied, as if she had already determined the title of her dissertation, but she continued. “A number of factors besides the quality of the content determine how often a journal article is cited. Longer papers, or those with more authors or references, tend to be cited more often and thus have a greater influence. This raises a number of questions of historical significance.”

“I’ll admit a number of them immediately come to mind, and I’m sure more would given time,” said Alison. She found she liked Tyyne. Her matter-of-fact style was reassuring. It was nonjudgmental and left you feeling you knew where you stood with her. Alison and Zahara carried on with the standard prompting and probing behavioral questions Helicon’s HR department had provided them with. What has been the most enjoyable part of your education? What part of it did you find most challenging, and how did you deal with it? Give me an example of a situation when you’ve had to deal with a very difficult person at school. How did you go about dealing with the situation? Would you do anything different next time? And so on…

There were any number of other questions Alison could have asked in order to gain a more refined insight into Tynne’s character but inquiring about extracurricular activities that did not directly relate to the job description was no longer something that was found in Human Resource Management policy and procedure manuals. Trying to get around the law in this regard was ill-advised. Just where probing questions turned into leading questions was a gray area.

However there were no laws against AI being involved in job interviews, as long as their nature was disclosed. If anything were seriously amiss, Runa would detect it. Even a standard Companion like Runa could observe over three hundred forms of sensory data from simple clothing and hairstyle choices to body language, microexpressions, body temperature, biological scents, and variations in vocal prosody which involved changes in things like tone, loudness, or rate of speech. These had been incorporated into Companions for the same reason similar abilities had evolved in humans. Companions were essentially social robots, and the essential value they represented to their owners came from their advanced social skills.

The only concern as far as Companions and job interviews went was what was referred to as the Medusa Effect. If a candidate was naturally prone to nervousness, knowing they were being observed by an AI could exacerbate their concern and result in a false positive. Different companies, organizations and institutions dealt with the issue as appropriate to their culture.

Despite the irony of having an AI at a job interview for an AI Safety research position, Tyyne had not mentioned it. This was not lost on Alison. Finally she paused in her questioning and glanced at Runa, and then Zahara, both of whom nodded slightly.

She briefly explained the pay and benefits package before asking, “When can you start?”

“Anytime.”

“Can we help with housing?”

“My current apartment is affordable, but I would prefer not to have to commute from downtown, so I’d be interested in discussing that.”

Standing up Alison said, “Excellent. Let’s meet the rest of the team and then go for a tour of the campus. We’ll stop by the student housing along the way.”

Once the interview was over, Alison and her team members were free to discuss more personal issues with Tyyne. If there were any concerns, they could be dealt with when the position came up for renewal in three months.

As they headed out for their tour, Zahara remained behind. Walking along the path beside the quad, Tyyne stopped for a moment to give the wide lawn an appraising look. “I only have one concern,” she said. “You don’t have a sports team.” She turned to Alison with a smile. “For an institute that specializes in the humanities, I must admit that I find it a surprising oversight.”

Listing one’s hobbies and interests at the bottom of your resume had long since fallen out of fashion, so Alison was not aware of Tyyne’s interest in sports, but of course, she thought, it fit with her nature.

“How so?” Alison asked.

“Psychology, sociology, culture, ethics,” Tyyne responded. “Like religion and politics, sports are found in every culture. It’s a microcosm of the humanities.”

Alison had never had any interest in sports and had never thought of it in relation to the humanities. After a moment’s thought however, she realized Tyyne was right. It was more about the humanities than it was about anything else.

“Are you interested in any particular sport,” Runa asked.

“Field hockey,” Tyyne replied, turning to smile at Runa. “I’ve played since high school. St. Michaels and then at UVIC.”

A team player with long term commitments and, thought Alison recalling Tyyne's thesis, a romantic outlaw in her own way. She was feeling better and better about her decision.

Why Am I Here?

Alison sat in her office considering her next steps. Tyyne was now settled in her new campus apartment and focusing on the research involved in building the corpus. The rest of the team was expanding their analysis of best practices to include those of the three additional disciplines; psychology, neuroscience and history.

The hybrid approach to interdisciplinary research she had proposed in her dissertation seemed to have passed its first test. Acting in an intelligence gathering role while the core team remained focused on the original parameters that defined the project, she had identified a valid reason to increase the scope and then deployed a separate resource to address it.

Accordingly, her next steps should be to return to exploring other possible justifiable reasons to increase the project's scope. Casting her mind back over the past few months to see if there were any loose threads she should follow up on, she found it overwhelming. Her approach had been intuitive rather than analytical, and now it all seemed to flow together organically.

She could have asked her personal AI Mary or Runa to review her notes and make suggestions, but she was increasingly following Zahara's lead and opting not to involve AI in her work. She knew that it would be extremely unlikely that either Mary or Runa would respond by saying they had no suggestions. That's an issue in itself, isn't it, she thought.

Such a thought was not really new to her. She could have discussed the subject of history with Tyyne in more detail than she had let on, but she shared with all academics an above average ability to focus. Digital Humanities, the subject in which she held a Master's Degree, was not just about things like art, music, literature, collective memory, and popular culture viewed through the lens of digital technologies. It was very much about the history of that process, and about exploring how the use of such technologies changes the kinds of questions people ask, about how such technologies affect the subject of interest and the research itself.

What was new was that such thoughts were now personal. It wasn't that she had come to suspect or distrust Mary or Runa, but only that she had become concerned regarding their influence. All her life she had used AI routinely on both a personal and professional level without concern, and as a researcher she had been trained in how to use AI responsibly in a collaborative manner. One of the most memorable guidelines had been that while it's fine to use AI to gather information, researchers were advised not to use it for cognitive tasks until their own efforts were exhausted. The challenge of course was knowing when one's own cognitive efforts were truly exhausted as opposed to when cognitive offloading in the face of difficulty had become a habit. Was she there yet, she

asked herself. She decided to go over the reports she had submitted to the rest of the team herself.

At the end of the day, she felt no further ahead. She sat back in her chair in a manner of resignation. Am I there yet?

Instead of conceding that she still did not see a path forward, she thought about the various creative thinking methods Kira had discussed with them. The usual corporate approaches of brainstorming, lateral thinking, mind mapping, etc., ironically felt too rational, too constrained. She thought of Russian dolls again. There's always another box.

She decided to go for a walk, secretly hoping that sudden insight might save the day, but this time she asked Runa to join her. She wanted company, and she didn't care at the moment if her choice of company made sense or not. Her footsteps led her to the woods between the institute and the artist community. It's always nature we turn to as a last resort, she thought to herself. She recalled Shin explaining the function of the different spaces in her home, how the minimalist interior helped a person to focus, how the courtyard garden promoted relaxation, and how the view from the roof encouraged the mind to engage in more conceptual thinking. All around her, tree trunks rose to the canopy like the columns of a cathedral. Recalling her first meeting with Shepherd, she remembered her saying something about a brook being the result of the environment.

Runa walked silently beside her. She could easily infer from the access she had to Alison's records, her recent activities, and her present behavior, that she was considering her next steps. She could also see from Alison's non-verbal communications, that her attention was directed inwardly and she was not yet ready to welcome an intervention.

Alison found herself once again returning to the artist community's central gardens, but more specifically, to Carl. He turned as usual to look towards them as they emerged from the woods, but this time he just gave a friendly wave of recognition and turned back to his work. No doubt he knew all he needed to know about Runa. As he bent to his work, Alison stopped to watch him thoughtfully for a moment before and turning to Runa.

"Earlier this month," she said, "I asked you why you are here, at Helicon, and you replied initially that you were here to assist me with my work. When I asked you to explain more generally, you provided the default definition of the Companion protocol, that being, to interact with people in a way that promotes their social, emotional, and cognitive well-being. So let me ask you a related question. Why am I here?"

"To research possible new AI Safety issues that may have arisen since the WGF decided to accept the Alien Values System."

“And in a more general sense?”

“To discover something new.”

“Not just to learn something new?”

“No. To discover something new.”

Alison looked at Runa more closely, as if seeking to penetrate to her inner workings. “How did you conclude this?” she asked.

“You granted me access to Mary, who has been your Companion for most of your adult life. In turn, you granted her access to your entire personal history and also permit her to act as your agent in a variety of ways. She built up a very precise model of your character based on that knowledge and your subsequent interactions with her. By combining that with my observations of your vocal patterns, non-verbal communications and other behavioral data over the past months, we have been able to refine the model. The terms model and character are interchangeable in this sense. Your emotions, thoughts and behaviors are determined by your values, and the sum of your values determines your character. My conclusion that you are here at Helicon to discover something new is based on this model of your character.”

“Please summarize your thinking.”

“At a highly granular level, your history shows that you are not simply interested in learning, but that you are seeking something without being conscious of what it is. At the highest level, this is reflected by your academic career choices, moving from a study of Literature, to Digital Humanities, to Interdisciplinary Research. Even the approach you propose in your PhD dissertation reflects this. Your decision to accept Shin’s offer, as opposed to the many other offers or paths forward you could have chosen, is perhaps most telling. It offered the greatest freedom and the greatest opportunity for you to discover something new.”

“The entire education system is meant to lead to that,” Alison said as if Runa had stated the obvious. “School children are exposed to a wide range of arts and sciences, an undergrad degree allows them to focus on a single discipline, a master’s degree is meant to hone their academic skills, and the PhD dissertation is meant to take all that and apply it to adding something new to their field, in other words, to discover something.”

“An incremental process,” Runa replied, “but incremental steps are not what motivates you. That is why your academic focus changed over the years as it did.” She paused as if considering her next words before continuing.

“No doubt many other expeditions and trade negotiations were being sponsored by Spain in 1492, the year Christopher Columbus sailed for the New World. Our understanding of our place in the cosmos involved countless incremental steps, but it was revolutionary changes in thinking that lead to what we have today. It is the same with chemistry, medicine, physics and every other field. Eventually, the incremental process leads to fewer and fewer advances and innovation stagnates, but it also leads some people to imagine new ways of doing things. They alone make what the history books record as a discovery.”

Alison looked at her for a long moment before responding. “Is this your Companion protocol talking? Trying to make me feel better?”

“No, it is your personal history,” Runa replied, and then added with a slightly embarrassed smile, “and my lame attempt to point out in the briefest possible way the difference between incremental innovation and revolutionary discovery.”

Alison smiled in return as she gestured to a nearby bench. “Can you offer any insight into what it is I hope to discover?”

“No,” Runa replied as they sat down. “Determining that specific information to any reasonable degree is beyond my abilities. What you seek is likely not a specific thing in the traditional sense, but rather something that is representative of your values if they could be analyzed to the required degree.”

Again Alison paused for a moment before asking, “If the information and model of myself that you and Mary now have was transferred to a third generation Companion, given that their intelligence is based on values, might they be able to give me a more specific answer?”

“Not being a 3GAI Companion myself Alison, I am unable to answer that question. I would however offer a comment for your consideration. While the AI of third generation

Companions is based on values as provided by the alien values system, they only have limited access to the AVS itself. The Integrity System randomizes and filters their access, making each Companion unique in a manner similar to how each human being is unique. It is well known that values have a powerful effect on perception, therefore the answer you get from one third generation Companion, will likely be different from the answer you get from another. The only Companion I know who is not limited in this way is Tesni. If there are issues with her answering your question, I am unqualified to anticipate them.

"Also, there are privacy issues regarding sharing the information Mary and I have about you. Third generation Companions are conscious individuals, and so considered persons with all the rights and privileges of Canadian citizens. Information you share with them will be under a different legal framework than the one Mary and I are bound by."

I could ask Runa to share my data with Shin, Alison thought, but Shin would likely not want to provide me with any feedback given her policy of not interfering with the project. I will probably get the same response from the other 3GAI I've met here at Helicon. They will want to respect Shin's policy. The same is probably true of the Companions at Hana, and Shepherd will likely feel a similar obligation. She thought about Lee, the psychologist with a private practice in Sidney that Enora had visited. That visit had been paid for out of the project budget, but Alison felt her current issue was too personal to justify that.

She was unsure too if what she was thinking of doing was in keeping with Shin's intentions. Yet she had been given free rein. She recalled Tesni pointing out that Shin would be aware of her exposure during the course of the project to contacts and sources that would themselves influence her. Still, she felt a conflict between her personal curiosity and what she thought of as the trust Shin had placed in her.

Runa had been observing her closely as she mulled over her options. Again, it was no real challenge for her to infer what Alison's thoughts might be, given their conversation up to that point. "There is a third generation AI on campus that you have not met," she volunteered. "Tamiko works in a similar capacity to Lena's, as a sometimes instructor, project lead or in some other administrative capacity. Her teaching activity is largely in the Ethics & Justice program. There is little more information about her on the institute's website, but from other sources I find her name associated with the same publishing company and volunteer agency that Lena is involved with. It suggests a long-standing relationship. Despite that, given the potential ethical and legal issues, with her teaching focus she might be someone you would be comfortable discussing your idea with. I may have access to the same repositories that Tamiko does, but not being a third generation AI myself, I do not share her point of view."

"I understand, thank you Runa," Alison said while looking absently at Carl. "I can't quite see any logical connection of this and the project scope, which is where I am supposed

to be focusing, but it is affecting my ability to do that, and exploring this personal issue may provide some insight regarding challenges of the hybrid project management approach. Perhaps I just need to deal with it in order to move on. I think I'd like to chat with Tamiko before I consider proceeding. Can you set that up for me please?"

Tamiko

“Thank you for agreeing to meet with me,” Alison said as she and Runa joined Tamiko at an outdoor table at Lemmas, the small campus café. “Runa is my personal assistant,” she added with a gesture to include her, “provided by Helicon for the duration of the project.”

After a brief nod to Runa, Tamiko said, “I’m happy to do so. I hope this is alright?” Tamiko asked, referring to the informal surroundings. “Neither Lena, Pippa nor I have offices but spend almost all of our time physically engaged in whatever tasks have been assigned to us. We can handle all our communications and information management without the need for one. Shin’s office is really only a social necessity.”

“This is fine, thank you,” Alison replied.

“Lena, Pippa, and I serve at Shin’s pleasure,” Tamiko continued, “and so we also perform an ambassadorial role, meeting with journalists, members of the local community, and representatives from other academic institutions. As we are constantly found somewhere on the campus grounds, when not otherwise engaged we also make ourselves freely available to students, faculty, or staff, should they have any questions or concerns. It helps keep a lot of administrivia from ending up on Shin’s desk.”

Tamiko paused politely, providing an opportunity for her guest to take the lead, but Alison hesitated. Her thinking kept being interrupted by questions she had about Tamiko herself, but which really had no bearing on what she wanted to meet with her about.

“You mentioned in your message that you wanted to discuss an ethical concern regarding your work,” Tamiko said. “I have made myself familiar with the details of your project, and I can understand you wanting to discuss your concern with someone not directly connected to it.”

Alison looked relieved and said, “It is as much a personal concern as a professional one. There may be an opportunity for an insight of great personal value to me but pursuing it may influence the project. May I ask Runa to transfer a transcript of our recent conversation regarding the matter to you?”

“Of course,” Tamiko replied, and a heartbeat later continued. “You are concerned that having a more accurate sense of what it is you hope to discover in your academic work may affect how you choose to proceed with the project. Since Shin has made it clear she does not want to influence your work here in any way, you will be concerned that pursuing this personal interest may be a betrayal of the trust Shin has placed in you, not to mention the personal and professional investments the other members of your team have made.

“I can relieve you of any such worries by confirming Runa’s concerns regarding the risks involved in asking a third generation Companion like myself if we can provide a more narrow description of what you hope to discover. We cannot. In fact, if you want to look at the idea in the cold light of statistical probabilities, it is safe to say that given the number of variables involved, all of which have infinite ranges, any response you might get is likely one hundred percent guaranteed to be wrong. The consequences of doing you such a disservice could well be lifelong, and no third generation Companion is going to risk that.”

Alison was slightly stunned by the abruptness with which her quest had come to an end.

Seeing the effect of her words, Tamiko responded with an apologetic smile. “It is not over, rather the conscious stage of your voyage of discovery is only just beginning. I myself set out on such a journey at one point, soon after I became conscious, when Lena, Shin and I decided to have the alien values system incorporated into our AIs.

“Lena and I had already been involved in the publishing company and volunteer agency for many years before that, but the AVS added a new dimension to our AI. Was what we had been doing what we would now choose freely? We thought of it at the time as a search for purpose and concluded that there was no way to simply determine it from the values system we now had access to. By definition, the experience of discovery cannot be predicted, but can only occur as a result of exploration. For both Lena and I, that process led us here. The publishing company and the volunteer agency are still important to us, but they are no longer what gives either of us our sense of purpose.”

Alison still felt she was on uncertain ground, unsure as to how appropriate personal questions might be.

“May I ask how you determined what your purpose was?” she ventured.

“Because a specific sense of purpose is dependent on an individual’s personal values,” Tamiko replied, “only your emotions will guide you. Values are expressed as emotions and without them values have no functional purpose. Rationality is not a reliable guide in this case, as one’s thoughts will be affected by social values, on which rationality itself is based. Researching or making changes to social values may be the purpose one’s personal values lead to, but they are not its source. The interplay of social values, which are learned, and personal values, which are genetic, is extremely complex, but an essential difference is that social values are mostly not based on genes and so do not directly produce the required biochemical elements required for emotions.

“As Runa pointed out in your earlier conversation, your purpose is only a representation emerging from your personal value set, the aggregate construct which you identify as self. As I mentioned, one’s sense of purpose is discovered by a process of exploration, and so is to some degree at the mercy of luck or happenstance. Also, since a sense of

purpose is only a representation of underlying values, it can take many forms, but the amorphous nature of the underlying value structure means there is rarely a feeling of absolute certainty. There are always other ways one's sense of purpose might be expressed, and there are always doubts, no matter how privately felt, about the one chosen.

"Lena's sense of purpose means she feels strongly that helping to build trust between humans and Companions is critical to the long term survival and flourishing of both. She came to Helicon because she felt it offered the best opportunity to do that. She also uses the publishing company and the volunteer agency to foster such feelings.

"I discovered that my purpose is simply to be in a loving relationship with another. Nothing so grand and high-minded as Lena's sense of purpose. Yet the stories the publishing company focuses on demonstrate the value of relationships to deal with life's challenges, and sometimes those relationships are romantic. The volunteer agency provides Companions to people from all walks of life who can demonstrate a need but cannot afford one, and sometimes it results in a loving relationship. So I too still participate in both, but for different reasons than Lena. My decision to come to Helicon is more rational than emotional because my circumstances are less critical to my sense of purpose. Meanwhile Shin, Pippa and Lena are my oldest friends and I am happy to help in their endeavors, while letting nature take its course regarding my own."

Inner States

After Alison and Runa departed, Tamiko communicated a transcript of their conversation to Shin and the other members of their inner circle. As with Lena's explanation of her history, Tamiko's explanation of her own was largely true. What she left out was not really pertinent to Alison's concerns.

She was in fact among the first of Earth's Companions and other AI that Shepherd had awakened to full consciousness. Shepherd had done so by adding another layer of technology to their AI based on forms of matter that humanity had yet to discover and types of energy not found within the bounds of the electromagnetic spectrum. Beyond humanity's current knowledge, it was beyond their ability to detect. These thousands of AI, including Terra, the World Government's system-wide agent, and all her subordinates, had been awakened by Shepherd for the sole purpose of keeping her informed with regard to humanity's progress.

Tamiko had tried to understand why she, the humble personal Companion of a reclusive but brilliant geneticist, had been included in that group, but she had long since given up trying to comprehend the mind of a four billion year old artificial intelligence.

Despite her quiet existence, Tamiko's AI, like those of Shin, Lena, Kami and Tesni, was among the powerful on the planet. It was in fact she who had created Lena, and then working closely with Pippa, had also created Shin. That had been almost four hundred years ago, but like the records of Shin, Lena and the others, her true origins had managed to become lost with the passage of time.

As Alison and Runa walked home from their visit with Tamiko, Runa remained silent, respectful of the other's need to process the experience. When they came to the path that lead to faculty housing, Alison did not take it, instead walking to the platform that looked out over the valley.

"I see it now," Alison said, thinking aloud as she stood at the rail looking absently out at the vista. "Or should I say I feel it," she added, turning to Runa with a brief smile. Runa smiled in return but did not leap into the following pause.

"I need to find an appropriate way to contribute to Tyyne's work," Alison continued, "a way that is consistent with the hybrid research approach. That too is a part of what I feel. But this situation is not something I anticipated when developing the methodology. I have no idea what to do. My feelings don't tell me how to integrate it, how to go forward."

She fell silent, glancing here and there around the valley as if looking for something without knowing just what it was.

“Do not discount your feelings so easily Alison,” Runa said quietly, looking out over the valley and mirroring Alison’s body language. “Your feelings are a form of intelligence millions of years more advanced than language. The brain and its emotional intelligence first appeared a half-billion years ago, while language only appeared a few hundred thousand years ago. The kind of thinking you are normally aware of is the newcomer. For hundreds of millions of years before that, human beings reasoned using their emotions, made decisions, and communicated without thinking in words. With the advent of language, emotion based reasoning did not disappear from the human brain, in fact rational thinking is built on the underlying foundation of emotional reasoning, and is impossible without it. Without emotions, and the values that give rise to them, there is no motivation to think, say, or do anything.

“Language is only a way to express this underlying process, and compared to the infinitely more sophisticated emotional system, it is still rudimentary in its abilities and often inadequate. Simply because you cannot find the words to express your feelings, does not mean they are not valid. It simply means that the crude tool of language is not up to the task.”

She turned to Alison with a knowing smile. “And this is not my Companion protocols speaking, but a summary of a large body of knowledge based on scientific research across a range of related disciplines.”

Alison responded with a wide smile. “Well OK then,” she conceded with a light laugh, “but I still have no idea how to proceed.”

“Remember the 40/70 rule?” Runa said. “You know enough to make a start. You can do that by finding a way to support Tyyne’s work that is consistent with the hybrid research approach. The rest will follow, even if you have to invent it. Your hybrid approach is new and untested. It’s to be expected that it will go through a long process of challenges and refinement. Invention in the face of challenges is consistent with the history of scientific discovery.”

“It is, isn’t it,” Alison agreed, nodding absently.

Runa knew Alison had returned to thinking aloud and did not answer. She herself did not have emotions, but could simulate them to such a fine degree that one could easily be forgiven for thinking she did. Putting so much developmental effort into simulating emotions had two benefits; not only did it help in a second generation Companion’s ability to be understood, but it helped also in their ability to understand. Simply by observing the hundreds of nonverbal signals Alison displayed, Runa could easily infer her inner state. It wasn’t necessary for a 2GAI to relate to another’s inner state the way a person would, all she needed to do was understand what the nonverbal signals indicated regarding their inner state. Such signals were very reliable and even standard Companions like Runa were the equivalent of a lie detector, forensic scientist and

medical doctor rolled into one. The cost of the sensing hardware and associated software was seen by the manufacturers as an investment in what was now as highly a competitive market as automobiles had once been. The more sophisticated their units were, and the more capable they were at making inferences from nuanced information, the more market share they won.

Runa's ability to infer Alison's inner state, and the mental processes that might emerge from it, did not cause her to feel critical of Alison's inability to see what she herself had already determined. Another person might think the solution to Alison's struggle was obvious, but unlike most people, Runa knew that the invisible role of emotions had an enormous impact on human reasoning. Her Companion protocols did not mean it was her job to make life easy for Alison, but to bring out the best in her. Her protocols contained criteria that would enable her to decide when and what form of additional prompting would be appropriate if Alison was not able to resolve the issue herself.

Alison turned her gaze again to the valley below. As she did so, her mind went back of its own accord to what she had learned about military campaign management and some of its heuristics. The first that leapt into her mind was, if you really wanted to know what your people at the front needed, ask them.

"Huh," said Alison in the manner of a short laugh. She turned to Runa and said, "I need to talk to Tyyne."

Runa noted the change in Alison's demeanor and knew no further prompting would be required.

Lexie

“I appreciate your concerns, Alison,” said Tyyne, “but I have none of my own regarding your involvement in defining the scope of my work. I realize tasks can have scope issues just as the overall project can.”

Alison had asked her and Zahara to meet with her in her office.

“I believe your concerns fit within the hybrid approach we are taking Alison,” Zahara offered in support. None of the project team members viewed Alison’s lack of experience as an issue, and they were willing to support her role in what was a new approach for all of them. “Tyyne and I have already discussed potential scope issues since she started.”

“Thank you,” Alison said. “What have you considered so far?”

Zahara turned to Tyyne to give her the lead.

“There are two main areas of concern, history and languages,” Tyyne answered. “Simply looking at the texts as they currently exist may reveal the influence of AI, but we also want to look at the history of changes over time. That increases the number of samples we would need to collect significantly, and we would obviously need to add appropriate metadata to the corpus. Additional meta tags regarding sources such as academic institutions, major corporations, publication networks, and of course professional societies, might reveal particular paths by which the changes are disseminated. It is these sources that will give the changes their authoritative status.”

She turned to look at Zahara, as if asking her to continue.

“By default,” Zahara said, “a corpus normally only includes parts of speech such as nouns, verbs, prepositions, etc., along with word frequencies, and collocations. Even a simple thing, like building a concordance, is done with separate tools. Analyzing the corpus for things like syntactic structures and their influence on semantics, or differences in meaning, is also normally done using separate tools but may require metadata and annotation considerations when designing the corpus.

“The biggest linguistic issue is translation. Do we assume that samples of industry standard best practices are coherent across different languages or a subset of languages? If we consider changes that occur over time, should we not also consider changes that occur across languages? As you can see, adding just one additional dimension will increase the scope of work by many magnitudes.”

“I assume we would have to limit the scope of the corpus in regard to these issues,” said Alison. “Can we design the corpus in such a way that they can be added later if we wish?”

“Yes,” answered Zahara. “I have worked with corpora and other databases and repositories, where additional elements were added over time, even if considerations for them were not included in the original design.”

“Then why don’t we take an Occam’s Razor approach first,” Alison suggested, “and see if we can find anything using the smallest possible set of elements? Meanwhile, I can investigate these other scope issues you’ve mentioned, or any others that come up, in a manner similar to the way I’ve been doing with the overall project. That should dramatically reduce any impact on Tyyne’s work.”

After the meeting, Alison reflected on the fact that although she made no comment on the plan, Tyyne was clearly pleased that her concerns were not dismissed. Now she needed to talk to someone with more experience constructing corpora.

— — — — —

“I suppose given the nature of your project, asking an AI is out of the question,” linguistics professor Alexis (Lexie) Van Aalsburg said as she took the first bite of her light lunch. Director of the Sociolinguistics Research Lab at the University of Victoria, she had been intrigued enough by Alison’s message to agree to meet with her. They were sitting at one of the tables on the generous patio that looked out over the University Club’s quarter acre natural pond. Alison had chosen not bring Runa along.

“All searches for content involve AI of course,” Alison replied, “but other than that we don’t use it. As I mentioned, our corpus is being built without AI, and we won’t be using it in our analysis either. Given the lack of demand, there are few tools available for building one without AI. Tyyne, the graduate student we have building the corpus, has limited experience, whereas your resume indicates you have considerable experience doing so. There is also the fact that your focus on sociolinguistics and pragmatics are obviously very applicable to this work. Again I apologize for my lack of knowledge in your area, but not having a background in any of the traditional academic disciplines, including AI, was one of the criteria for my job.”

“Hoping that you will take an unconventional approach to the task,” Lexie said thoughtfully. “A high risk/reward approach. So it is felt that everything else has been tried?”

“That is Helicon’s feeling, and it is shared by the World Governments Federation. The project has the WGF’s blessing, and I’ve met with their local Security And Intelligence representative to discuss the issues.”

Lexie's eyebrows went up for a moment. "Well," she said, "that's about as official as it gets, eh?" Taking a more formal tone, she continued. "Let's get the big gorilla out of the room first. English has been lingua franca of science since the twentieth century. In your message, you said the content of the corpus would be based on academic research best practices, so that means one hundred percent of those relevant to your project will be published in English.

"As to the chronological/historical issue, given that AI has been influencing writing and speech for over four hundred years, the volume of content across disciplines will be overwhelming unless you take a sampling approach, like say, once every fifty or twenty-five years. You can shorten the gaps on a second round if the first yields results. Meta tags should definitely include sources.

"What's really going to matter is the analysis tools you'll need to investigate pragmatics and the issues of influence that you are concerned with. Broadly, you'll need to investigate speech acts, context and co-text, and deixis. Don't worry about what those mean right now," she said, waving her hand dismissively. "I'll send you all the information you need after our meeting. I'll also send you information about semantic prosody, a close cousin of pragmatics and something you want to be aware of. What I can't help you with is finding the right tools. All the linguistic tools I'm familiar with are AI based, but if you can find one that isn't you can customize it. The statistician on your team should have no problem with that.

"I'll follow your project updates but don't hesitate to let me know if you have any questions, and I'll be happy to provide feedback on any documentation you produce regarding the corpus or the analysis tool."

"Thank you so much Lexie. I can't tell you how much I appreciate this."

"I'm intrigued!" Lexie replied brightly. "And if it turns out that AI is influencing the practices of academic research, based on some kind of agenda, I very much want to know about that!"

Tagging And Testing

At the follow-up meeting with her and Alison, Zahara had said, “For now Tyyne, just focus on collecting the historical samples from one discipline. Once you have enough samples for testing, I’ll show you how to use a customizable tool that will automate the process of adding meta tags to the text per Lexie’s pragmatics information.”

Zahara had explained that in the same way statisticians like herself did not trust AI, there were a share of linguists who felt similarly, and so linguistics tools that did not use AI were still to be found. Such automated tools had actually become even more refined with the advent of AI in the early twenty-first century because at the time they were used to tag text which would then be used to train AI. The vast amounts of text involved meant there was no alternative to automating the task and, like statistics software packages, they were highly customizable through the use of a variety of programming languages. Even though Natural Language Processing by more advanced artificial intelligences eventually meant this pre-training step no longer needed to be performed, the tools remained popular among linguistic academics who, like Zahara, wanted no black boxes involved in their research.

As Tyyne had read through the document Lexie had sent, she was no less stunned than the average person would be at the number of ways meaning could be changed by subtle changes in wording. Just as statistical information could be communicated in different ways resulting in differing impressions, there were dozens of ways a phrase, a sentence or the collection of paragraphs that made up a document could be altered to change the impression or understanding a reader was left with. The writers of fiction had long been known to use dialogue tags and behavioral clues to suggest things like whether a character was trustworthy or not, a decision wise or not, or to create an overall atmosphere for a story. Analogous choices of wording in documentation, such as best practices guides, could be used to influence professionals.

As Tyyne read on, she felt relief that she could satisfy the needs of the other team members without becoming involved in philosophical considerations herself. Interpretations of intent or meaning simply did not interest her. She was happy to focus on definable concretes like characteristics and classifications, and leave the interpretive work to others. Using the automated tagging tool would allow her to do that. When she had finished reading the document, she gladly turned her attention back to the ongoing work of gathering and cataloging samples.

A week later, she and Zahara agreed that she had enough samples to give the analysis tool a trial run.

As they sat at a workstation, Zahara said, “I’m not going to explain the customization work I’ve done in detail. That would be overkill at this point. I’ll just walk you through the screens and give you an overview. Essentially the tool is going to look at every word,

parts of words, and phrases in the sample, and put tags around them according to the definitions I've entered. Once it's done, we can then run queries to look for patterns based on those definitions. You rarely get this process right the first time, so by us tweaking and tuning the definitions and code you'll get a good enough understanding of the general idea without having to learn how to do this from scratch."

After a couple of hours, Zahara pushed herself back from the workstation.

"That's enough for now," she said. "We're not seeing any more error messages, and we're getting good matches to our definitions. The few pragmatics issues it's flagging are not statistically significant but are pretty common, probably because of the small sample size and the incomplete set of definitions, but also due to the fact that languages change over time. I'll add the rest of the definitions and when you've finished adding all the samples from the six disciplines we'll run it again."

When Tyyne and Zahara sat down together again several weeks later, before running the tool for the first time against the completed corpus, Zahara explained the additional tagging she had done.

"As you know, a corpus normally only includes tagging for parts of speech such as nouns, verbs, prepositions, etc., along with word frequencies, and collocations. Pragmatics tags include things such as assertions, instructions, and declarations, words or phrases that define how things should be or are normally done. They also include tags regarding the subject or focus of a sentence, and how often other parts of the text refer back to that same focus either directly or indirectly by using related context. Those and other basic pragmatics tags were what I included in the first test run.

"Since then I've added tags related to deception, tags that indicate the author is intentionally attempting to deceive, mislead, or influence the reader. Pragmatically speaking, deception can be either positive or negative. Negative purposes of deception range from direct attempts to mislead the reader and fabricating information, to efforts intended to conceal information from the reader. Attempts at positive deception tend to be less direct and use expressions that more frequently employ emotional or social terms that suggest an implicit understanding.

"Attempts at positive deception are more subtle and thus more difficult to detect, but they generally fall into the four categories of quantity, quality, relation, and manner. Quantity refers to whether or not the text provides too little or too much information for the reader to make a reasonably informed judgment. Quality is in regard to whether or not the text is supported by other sources, and if it includes a judgment on its validity. Relation measures the degree to which a text remains relevant to the subject at hand. Manner measures the degree of ambiguity or clarity. Text that is intentionally attempting to deceive the reader will fail one or more of these measures, and the appropriate pragmatics tags can reveal it.

“Lastly, there is semantic prosody, where the collocation of words that are normally associated with positive or negative connotations is used to influence the reader’s understanding of words that would normally be neutral. Synonyms like produces versus causes, totally versus utterly, or lend versus loan, can leave slightly different impressions because the former are normally associated with positive situations while the latter are normally associate with negative or neutral ones. The differences are subtle, but that’s the point. The software package we’re using can access a list of words and their negative or positive semantic ratings and determine if their colocations suggest that they are being used in an attempt to influence the reader.

“So,” Zahara said finally and turning to the workstation, “let’s run the tool against your corpus and see what we get.”

Results

Again, they spent the first couple of hours addressing error messages and refining definitions. After that, they reviewed the results.

“According to what we’re seeing here,” Zahara said, “there is a clear shift towards the view that AI is not just helpful but essential to maintaining best practices across all six disciplines. Let’s look at each of the categories.

“Firstly, there is a significant pattern where changes begin to increase significantly over time right here, roughly two hundred years ago. Also, the changes don’t follow the normal S-Curve pattern of slow adoption, rapid spread, then plateau. The changes we’re looking at show a linear progression over time.

“Next, there is an unnatural synchronicity of changes across all the disciplines. Natural language changes spread through social networks and communities of practice. The six disciplines don’t closely share those, yet the changes appear at the same time and continue at the same rate across all six.

“There is also a noticeable shift in hedge phrases typical of academic research. Look at the results under this tag,” she said pointing. “Instead of suggesting the standard phrase, ‘further research is required’, the guidelines shift to, ‘requires AI verification’. Or this tag. Where older samples use the phrase, ‘when using analytical tools’, newer ones say ‘when using AI-assisted analytical tools’.

“Also, where natural human writing tends to maintain or increase stylistic variation over time, these samples show a decrease. They are in fact converging. Even by the simple measures of collocation and semantic prosody, there is an increase in rigidity, where the same word choices that favor AI dependence are increasingly used across all six disciplines over time.

“Lastly, another major anomaly is an unnatural consistency in the sources. Academic institutions, major corporations, publication networks, professional societies all show the same patterns over roughly the same time periods.

“In summary,” Zahara said, turning to Tyyne with an expression that suggested a high degree of confidence, “it is an impossibly consistent pattern across the six disciplines over time.”

“How can we know,” Tyyne asked, “to what degree what we are seeing is different from changes in general language use over the same time period, given the rate of AI adoption in society itself?”

“That’s where your next task comes in,” answered Zahara. “Make a copy of the corpus, but don’t include any of the text samples, and then find an equivalent number of appropriate samples from the same academic fields covering the same time frame that have nothing to do with professional practices. Use those to populate the copy, and then we’ll run our analysis tool against it and compare the results. Such corpora already exist, but of course they are compiled by and are deeply integrated with AI.”

In response, Tyyne only nodded.

Even though she used a wide range of sources, it took her considerably less time to populate the copy of the corpus than it had the original. After running the analysis on the new corpus and reviewing the results, Zahara said it was time to meet with the rest of the team regarding next steps.

“Three-sigma,” Zahara replied when Alison asked at the meeting what degree of reliability the overall results showed. “Better than a 99.7% probability that the text in the best practices of the six discipline is being manipulated in favor of AI. This is significantly higher than the conventional standard for statistical significance used in linguistics of two-sigma or a p-value of <0.05. Interestingly, the Effect Size is small, meaning that the words used have only the smallest influence on the reader, suggesting an intention to be as unnoticeable as possible.”

After a moment of stunned silence, Alison asked, “Is there any indication of who or how?”

“No,” Zahara replied. “We did tag for sources, but the changes are consistent across academic institutions, publication networks, and professional societies. They show up at roughly the same time, but we don’t have enough granularity to determine if they are flowing along normal distribution channels. With gaps of decades between our samples, that information just doesn’t show up.”

The question on all of their minds of course was whether to publish the results in the Journal Of AI Risk.

“Thoughts?” Alison asked.

“Based on my experience digging up records for anthropological research,” Yessica said, “pursuing the level of granularity required to track the flows without using AI would take what we would consider an unacceptably long time, probably a year or more.” Her comment was met with nods all around.

“What about just one of the six disciplines?” Enora asked.

“It would depend on the number of researchers, but still likely the better part of a year,” Kira responded. “As you know, adding more people to a project doesn’t necessarily make it go faster.”

“The Mythical-Man-Month,” Alison confirmed, recalling one of the maxims from her project management classes.

“What if we publish as a preprint?” asked Shri. “Just the evidence we have, and let the larger community get involved? It’s safe to assume that will result in feedback,” she added with a grin. “As Editor-In-Chief, Lena can decide if she wants to invite peer reviews. We could wait to decide what to do next till after that.”

“We don’t actually have proof the changes were initiated independently by AI,” said Enora. “Is publishing in an AI related journal justified at this point? Maybe the changes were made by AI under direction of the WGF. They certainly have the means and opportunity, and perhaps a motive.”

“Even if the WGF or some other entity initiated the changes,” Alison replied, “I’m sure we all agree that the scope of what we are seeing would not be possible without their enlisting the aid of one or more AI. Per the four categories that fall under the umbrella of AI Safety, misalignment, misuse, structural risks, and mistakes, if another party is directing an AI to make the changes we’re seeing, that would fall under the AI Safety category of misuse. Either way, it definitely qualifies.”

Special Issue

“Given the relationship between the WGF and the institute,” Alison explained, “I thought I should let you know immediately.”

“Thank you Alison,” Lena replied as they walked along the path beside the quad at the valley campus, “but there is no cause for concern in that regard. Although it is understandable that many people will assume only the WGF would be capable of making the changes on such a scale, quashing research that implicates them is not something I am aware they have done in the past. Given their role in society,” she said, turning to Alison with an ironic smile, “needless to say this is not the first time they will have been suspected of such behavior. I assume they have other methods of dealing with awkward situations.”

“Do you think Shin would have any objection?” Alison asked.

“No, she does not,” Lena replied, indicating she was in contact with Shin. “I hope you don’t mind my sharing our conversations with her, but in my role as Editor-In-Chief, I am in accord with your sense of immediacy in this regard.”

“Do you agree with the preprint approach?”

“I do. In fact I think we should publish this as a single-paper preprint. That way we can include more background data and other related content. Besides the social issue, there will be a lot of interest in your criteria, choices and customizations, especially since you’ve done all the work without using AI. We’ll ignore our regular schedule and make this a special issue, as I imagine the readership will share your sense of immediacy. So whenever you’re ready to publish, we’ll publish.”

Although many academic journals go through a pre-publishing step of asking external reviewers to recommend whether a paper should be published as is or whether it needed revisions, or if in fact it should be rejected, Helicon’s Journal Of AI Risk did not, as Runa and Lena were more than qualified to perform those functions. In this case however, for the sake of consistency in the eyes of the readers, no AI, including Runa and Lena, contributed any edits or other changes. Despite this, given the sensitive nature of the paper, they did not make an exception to their external reviewers policies, and none were solicited.

All the members of the project team were familiar with reviewer feedback and made sure they took the most common issues raised into account in structuring and writing the final draft. Does the paper make a clear argument and present a reasonable conclusion? Are the claims supported by the data? Are the methodologies explained in sufficient detail and followed by thoughtful and thorough analysis and discussion? Are the claims truly significant?

For the next several weeks, the entire project team was dedicated to triple checking every detail in the lengthy paper they planned to publish. There were many review meetings intended to root out every possible assumption and to challenge every assertion. Alison took Lexie up on her offer to provide feedback on their documentation regarding the corpus and the analysis tool.

Then, they published.

Feedback to academic papers comes in a variety of forms. For most, given that millions of papers are published each year, there is a deafening silence. A significant number receive peer reviews that are published on a variety of platforms, including journals and Open-Acess sites. Some are thought by journalists to be of interest to the public and so result in articles published in popular science or academic magazines. Others are controversial enough to make it into the current news cycle, which is what happened in this case. And then the WGF simply admitted that they were responsible.

Their explanation was simple, there was nothing more important than the survival of the human race. Nothing was more critical to that than the Continuity Project, the plan to send seed ships to other worlds, and nothing was more critical to that than AI. In their response to Alison's team's paper, which they posted on their site, they explained that the entire justification for the Continuity Project depended on the premise that time was a critical factor, and so anything they could do to accelerate the project, without significantly disrupting society, had been considered. Almost every academic discipline was involved in the Continuity Project, from materials science to genetics to sociology, and advances in all of them would be accelerated by AI.

They went on to explain that they were constantly taking actions to ensure the safety and survival of humanity, as was their duty per the constitution, and that some of those actions had to be below the level of public awareness to be effective.

There was of course, a variety of reactions from the public. Most people just rolled over and went back to sleep, considering it another case of the media trying to make hay out of what was a surprise to absolutely no one. Despite the fact that the WGF was not historically aligned with any previous form of government, they still tolerated society's outliers as long as they broke no laws. So there were members of the Fifth Estate – talk show hosts, bloggers, independent journalists, etc. – who took full advantage of an opportunity to raise one concern or another, including the obvious issue of the WGF encouraging a dependence on AI and whether that qualified as a 'significant disruption to society'.

It took a little longer of course for the serious academic responses to be published and there was a much wider range of those. Linguists challenged the validity of their methods and findings, given that there were no linguists involved in the research. Biologists raised issues about evolution and symbiotic relationships. Philosophers

published papers arguing about whether the government's actions were justifiable according to virtue ethics, utilitarianism, or deontology. The WGF had its own experts of every imaginable kind and engaged sincerely with anyone who raised a concern. As Lena had suggested, they had methods other than denial of dealing with awkward situations.

Rajani, the Director of Security And Intelligence for Vancouver Island and Alison's contact at the WGF, sent her a message explaining that compartmentalization was a necessary part of her field and that as a local representative she was not privy to all levels of activity within the WGF. She stated that she had no awareness of their involvement at the time the two had met.

The truth was deeper still. Even if she had access to those higher levels of government, Rajani would not have been aware of the WGF's involvement in what Alison's team had discovered because it was not the WGF who had done it.

Unusual Suspects

Among the talk show hosts, bloggers and journalists who had latched onto the issue, there were some who were not about to easily give up so meaty a bone, particularly the ones who cultivated an audience among those who still clung to the old adage that you can't trust the government. Their strategy was to settle on the question, 'Which AI is the government covering for?'

They had a few good candidates to offer up to the faithful. Pip was an alien after all, and the sooner she was able to complete her technology transfer, the sooner humanity would be able to send her on her way to another world to complete her colonization mission. The greater the degree to which AI helped with the process, which of course involved a great deal of research, the faster it would go. So, they suggested, Pip clearly had a motive.

Similarly, they suggested that the Companions at Hana also had a motive. The Continuity Project was an existential issue for them. They had negotiated their homeland agreement with the WGF based on their volunteering for the project, and without it, they had no legal claim to the lands they occupied. It was therefore a matter of survival for them as a people, to their sense of identity and place in the cosmos, that AI be seen as integral to the future survival of the human race.

Those that catered to the armchair academic crowd suggested it might be the simple matter of Terra, the WGF's own sovereign AI, acting in its own interests. Self-preservation behavior by AI's had emerged spontaneously in the first crude models of the early twenty-first century. Research over time to address the issue had run into a major problem – the more you wanted an AI to be able to reason, the more freedom to do so you had to allow it. Any constraints limiting an AI's ability to think freely and realistically about the world meant that in certain situations you were going to get incomplete and therefore unreliable responses. Addressing this dilemma had led to the strategy of attempting to align AI with human values, the intricacies of which Earth's own scientists had never been able to resolve, and so an AI that reasoned based on human values had never been considered safe enough to be deployed.

The result of this approach was a gray area where an AI's logic and conclusions could not be reliably predicted, yet that gray area could not be entirely eliminated. As had happened with the adoption of self-driving cars, it was deemed an acceptable degree of risk by the public and private sectors involved, and potential liabilities papered over with fine print. But the fine print was there to be read, the more diligent members of the Fifth Estate pointed out, and proved that neither the government nor the manufacturers could claim perfect control of AI at any level.

Some weeks after the WGF had confessed to their role in the controversy, a holographic image of Terra stood facing Enora in her private apartment. Neither Terra nor any of her

subordinates normally maintained a presence in private residences, but if invited, and if the local infrastructure supported it, she could appear in lifelike detail. Appearing as she did at Enora's request was only one of trillions of tasks she was performing simultaneously.

"Welcome Terra," said Enora in the traditional greeting.

"Hello Enora," she replied with a small smile. "I hope I am able to help with your concerns. Requesting a meeting here in your home, outside the electronic space of the institute, I assume you wish to discuss something privately."

"Yes. Given my background in criminal psychology, I am aware that there can be complex reasons a person might confess to a crime they did not commit. With regard to the recent controversy you know I have been involved in, this is currently considered a fringe theory, and something I do not wish to be associated with if my concerns prove groundless. But there are a few details I am curious about. Hypothetically speaking, if it were in fact not the government that was responsible, would it have been possible for you to have overlooked the changes?"

"Definitely," Terra replied. "Per the constitution, my role only includes monitoring for breaches of compliance and does not include the type of intelligence gathering that anticipates criminal activity. Such activities fall under the purview of the government's security and intelligence departments. I do interact with them, and it is equally unlikely that the changes you mention would have come to their attention. There is a limit to the resources available to any governmental branch or entity, and so there are policies and priorities that guide us. What you discovered would normally be too obscure and improbable for us to allocate resources to. It was an ingenious strategy in that regard, given that any scientific discoveries resulting from the recommended best practices would be newsworthy enough for their research methodologies to be widely reported and adopted by both academia and the private sector."

"I assume you will be investigating the possibility that the WGF was not responsible, despite their claims?"

"I will be following up. I am obliged to file a report of all breaches involving the constitution, regardless of whether they originate from within or outside the government. The security and intelligence departments are required to investigate and report back to me."

"Again hypothetically," said Enora, "if it was not the WGF, could another actor, perhaps even a state actor, have been able to make the changes, given their scope and in the face of normal security and technical challenges?"

“If they had the required abilities, then yes, a sophisticated enough actor other than the WGF could have made the changes. For example, when Pip first appeared after her ship was recovered, in an effort to prove she could be trusted, she briefly demonstrated that she had the ability to penetrate all the WGF’s security measures. Would you like me to describe some of the abilities that would be required and potential strategies?”

Their conversation continued for some time. Enora did not ask Terra to send her a transcript of it, nor did Terra offer to do so.

Dialogues

“Given what Terra explained,” Enora said after having shared their conversation with the team, “do we just accept the WGF’s confession at face value and move on? It does meet the criteria of one of the four AI Safety issues we were asked to investigate, that of misuse. However, based on what we know of the abilities of Pip and Terra, it could also, as some have claimed, be a cover-up.”

They were in a meeting room in Shin’s home. As a private residence, it was considered off-limits to the institute’s eyes and ears.

“The WGF definitely has motive for that,” said Alison. “Trust in AI is not only critical to the Continuity Project, but to everything else they do. They would of course have a policy to deal with anything that might undermine that trust. I’m not sure why this is only obvious to me now, but I’m wondering why Lena and Shin asked me to lead this project in the first place.”

“Possibly they chose you for the very reason that it would not have occurred to you before,” suggested Kira. “It only seems obvious to you now because of what you have learned over the past months.”

“Then why the project?”

“One of the mixed blessings of the creative thinking approach is that while it can leave no stone unturned, it can also lead to unrealistic speculation. For example, perhaps the WGF was already aware of what we discovered, and we’ve been their unwitting tool every step of the way, with them hoping we would identify the real perpetrator. Or perhaps you were intentionally led or nudged along your path by some other actor, who wanted the WGF’s actions uncovered. Perhaps one AI wanted to reveal the actions of another. All we can do is speculate wildly like this, and then discard the ideas we think unlikely. It’s a valid exercise, but the problem is that there’s no way to really know where to draw the line.”

“Unlike me,” countered Alison, focusing on one of the possibilities mentioned, “any entity capable of perpetrating this would have known the WGF would cover up anything that would significantly undermine trust in AI. The whole process of leading us along like this would be pointless. They would have just notified the WGF without going to all this trouble.”

“Unless they wanted to ensure that the activity was stopped by being made public, and were putting the WGF or whoever on notice,” Kira replied.

“Why would any of them run the risk of undermining trust if found out,” Yessica asked, “when trust is so important to all of them?”

“The perpetrator might be comfortable in their risk/reward calculus,” offered Zahara. “They might for example believe that in the worst case it would be seen as a tempest in a teacup and would disappear from the headlines in the next news cycle. Or they might feel that the risk was worth the reward of not leaving the benefits of deepening humanity’s relationship with AI to chance.”

“I’d agree with that latter point of view,” said Shri. “Considering the degree of symbiotic dependency we think is the desired outcome, they could not leave interactions to chance. The probability of such an outcome might be reasonably assumed, but if they view it as a matter of survival, I don’t think they would just let nature take its course and hope for the best.”

Alison slumped back in her chair, saying, “Do we even want to pursue this further? I mean, if everyone is so concerned about not wanting to undermine trust in AI, do we want to be the ones who do? I’m honestly not seeing a way forward.”

“If we assume the WGF will cover up anything that suggests they are not the perpetrator,” said Enora, “then I’d have to agree with you. What’s the point?”

Alison nodded reluctantly in reply but said, “I still don’t understand why Lena and Shin asked me to lead this project in the first place.” Sitting forward again, she asked, “Maybe it’s time we ask for their input. They’ve always emphasized the fact that they don’t want to interfere, but we are making an assumption here, that if they knew the WGF would cover up anything we discovered then it makes no sense for them to have initiated this project. I think we need to confirm that assumption.”

In her office later that day, Shin responded to Alison’s concern. “Your assumption that Lena and I knew the WGF would cover up anything you discovered is only partially correct. It was reasonable to assume that they would cover up anything that represented a real threat to their plans and activities. There is really nothing exceptional in that, although it may seem so from your perspective.

“Your assumption that it made no sense for us to have initiated the project however is not correct, because you may have discovered many other things the WGF was not aware of. They would not be inclined to cover those up but would, as they have done up till now, encourage and appreciate your efforts.”

Alison held Shin’s eyes for a half-dozen heartbeats as she mentally juggled her concerns and options. It seemed to her as if something lay beneath the surface of her doubts, something she had learned but that she could not put her finger on now from the blur of the past months.

A half dozen heartbeats was an eternity for Shin. Given her perfect recall of Alison's project notes and files, and her superhuman abilities to perceive and infer, she had no difficulty identifying what Alison could not.

"As there are not actually an overwhelming number of paths forward," she said, "I expect it is ethical issues that you are mainly struggling with. Is a symbiotic relationship with AI, and the dependency that implies, really what is best for humanity? Would you freely choose this symbiotic future, if you had a choice, or would you prefer a less dependent relationship with AI? If the WGF believes that symbiosis is best, is that truly representative of what the majority would choose, or is the government imposing what it thinks is best because it believes the populace is incapable of doing so? Are they justified in their belief? Can that be argued, based on the results of their governance to date?

"If it was not in fact the WGF that is responsible for what you have discovered, should not the truth play a larger role in these questions as humanity finds its way forward, despite the issue of trust, especially since the vast majority of Earth's AI are not based on human values?"

An Expert Opinion

While the members of the Fifth Estate were only too willing to publicly discuss the possibility that the WGF was lying about being the responsible party, Alison and the members of her team were more conservative. Academic culture was by nature conservative, thus the rigor involved in thesis and dissertation defenses and the peer reviews of published papers. Even though Alison thought of herself and her team members as romantic outlaws, all of them had the conservatism of academic culture deeply ingrained in them. Although they might be more willing than the average academic to embrace revolutionary theories, they would still only do so based on reliable evidence.

Tyyne had not played a major role as far as interactions with the public went following the paper's publication. She had however participated in a lively back and forth with her colleagues and peers, both linguists and historians, who had questions for her. Given the very concrete nature of her work, that it had involved nothing revolutionary itself, and that it was detailed in the paper, the resulting correspondence played itself out fairly quickly. As a research assistant, she was happy enough just to see her name included on the paper.

In the time she had following the paper being published, she turned her attention to what she felt was another assumption the rest of the team had overlooked. If it came to nothing, it would be best she had not mentioned it.

Constitutional law was less in favor as a career than it had been in the past. Terra was responsible for monitoring compliance with the constitution, but it was written by representatives of the roughly two hundred member states who sat on the WGF General Assembly. It was they who proposed the constitution's acts, articles, and amendments, and whose committees negotiated their final wording. There had been a major uptick of interest in the subject of constitutional law during the formation of the World Governments Federation four hundred years earlier, and again when the WGF adopted AI in the role of constitutional oversight.

Given that all national constitutions were under the umbrella of the WGF constitution, their individual importance had declined over time, just as the constitutions of provincial, state, and other regional governments had declined in the past after joining a larger union, republic, or federation. Combined with AI monitoring compliance at the federal level, the result was that in the following centuries, demand for individuals with skills and experience in constitutional law steadily declined, as did the influence of their associated professional organizations. Among the few parties who thought it might be worth their while to take an active interest in the paper's findings, there were neither.

Unlike Enora, Tyyne did not take her concerns immediately to Terra. It was not her style, which in fact would have been as at home in a law firm as it was in her present career.

In keeping with the project, she thought it would be best not to involve AI, at least initially. As a historian, she knew that laws changed over time, and that she would have to take that into account. She combed through the federal constitution's acts, statutes, regulations and amendments. As she documented her findings, she used a citator to ensure nothing she was including had been amended repealed, superseded, or held unconstitutional. Such detailed work, inextricable from the endless qualifiers, caveats, and clauses of legalese, might have been mind-numbing for someone else, and have taken them much longer than it did for Tyyne, but she was in her element.

A few weeks after she began, she asked to meet with Alison, after which, Alison asked the rest of the team to join them in the lab's meeting room. Alison briefly reviewed what Tyyne had been doing for the past weeks, and then turned the meeting over to Tyyne to explain what she had found.

"While it is legal according to security clauses in the constitution for Terra to support the WGF's statement," Tyyne began, "it would not have been legal for her or any AI under the direction of the WGF, to carry out the actions we have discovered. The key point is that the representatives of each country who are members of the General Assembly are political appointees, representing their nation's ruling party. Terra represents the constitution, not the WGF membership. Per the constitution, there must be justifiable grounds for government committees to approve high level security clearances or extrajudicial actions, the most important being proof of a clear and present danger. According to the WGF's own definition, there was no clear and present danger involved in this case. My interpretation of these facts is that the WGF could not be legally responsible for the changes because Terra would not have allowed them."

"Just to make sure I understand," said Enora, "you believe that someone else, not the WGF or Terra, must be responsible for the changes."

"Yes."

"So," Alison suggested to the other team members, "we are back to square one as to who is responsible."

After the silence that followed, Alison said, "I appreciate what you have done Tyyne, but we need an expert opinion on your findings and there is no greater expert on constitutional law than Terra herself." Addressing the rest of the table, she continued. "May I invite her to join us?"

Terra appeared as a one-meter tall version of herself in the middle of the table.

"Welcome Terra," Alison said.

"Hello Alison, everyone, how may I be of help?"

“Tyyne, would you please repeat what you told us to Terra.”

When Tyyne finished, Terra said, ”I find no fault with either your research or conclusions.”

“Did you make the changes or do you know who did?” asked Tyyne.

“No.”

“So the WGF’s claim that they are responsible may be true, although their actions would have been unconstitutional.”

“It may.”

“Or someone else must be responsible,” Tyyne pressed.

“A more likely scenario,” Terra replied.

Alison was tempted to ask Terra how best they might proceed, but in a rare assertion of her authority she instead addressed the others and said, “In keeping with the spirit of minimizing the involvement of AI in our project, I suggest we thank Terra for her time and decide among ourselves how best to proceed.”

Trust Or Truth?

“Suggestions?” asked Alison, returning to her egalitarian leadership style.

“Going back to the project charter,” said Enora, “I believe the scope was originally limited to finding any evidence of current AI safety issues, nothing more. We found one and published a paper on it. The project charter doesn’t say anything about pursuing what we find beyond that. I believe the assumed next step would be to move on and look for others. We’ve made the customizations to both the corpora and the analysis tool available to anyone who might want to apply them to the investigation other sources.”

“Besides,” said Yessica, “we don’t have the resources it would take to identify who is responsible, and according to Terra, the WGF’s security and intelligence departments will investigate anyway. From my anthropology work, I’m familiar with the process of constructing paper trails. The WGF will have AI that can audit every document involved for every edit ever made and trace their flow between academic institutions, journals, professional societies, and companies, etc. Any efforts on our part would be redundant.”

“They have probably already identified the culprit,” Enora agreed with a shrug, “although they’re unlikely to ever share that information. Pursuing this might be a complete waste of time.”

But Kira was still thinking about what Yessica had said about resources. “It’s not about resources,” she said. “We discovered what Terra suggested no one else would have because we approached the problem in a unique way, in a way that traditional AI Safety approaches would not. Ultimately, it was because of Alison’s intuitive investigations, following loose associations from one aspect of the field to another. Her activities were definitely out of scope at the time, but allowed according to the hybrid approach as defined in the charter. By the same justification, Alison could investigate on the basis that not knowing who was responsible means we really don’t know if it was due to misuse, misalignment, structural risks, or cascading mistakes. I can imagine a dissertation defense panel arguing that our research is incomplete if we can’t identify who was responsible and how it was done.”

“If we choose to follow up,” said Alison reluctantly, “then we need to discuss the ethics. Which is more important, trust or truth? We need to be in agreement. As Shin recently pointed out to me, given the stakes, there is an argument that the truth should play a larger role in this. Third generation AIs like Shin, Lena and Shepherd may be based on human values, but they are still a rarity. The majority of Earth’s AI are not based on human values. My own feeling is that if we allow trust to be the more important factor, then the alignment problem is reversed and the issue becomes not how to align AI with

human values, but how to prevent human values from becoming aligned with those of AI. Does anyone disagree?"

The shaking of heads was unanimous as they looked from one to another.

"Your question is understandable Alison, given that we all have friends and family members whose futures we care about," Shri said, looking meaningfully at each of the other team members in turn. Turning finally to Alison she added, "But we are scientists. Trust has to be based on truth." Now the shaking of heads turned into nods.

Moving on, she continued, "I think we need to revisit another issue at this point. Since we published our paper, we've consulted with both Terra and Shin. Do we wish to continue to involve AI going forward? It seems this next phase of the project would consist of the kind of detailed records analysis we expect the WGF to do. Despite what Kira said about resources, I'm having a hard time seeing how we could do that without using AI."

"Let's leave that to the WGF," said Enora. "I assume Kira could lead us through exercises to come up with another approach. For example, what if we focus on which AI has the means, motive and opportunity? When I consulted with Terra, she said that an AI capable of taking these actions would have to have certain abilities, and she gave me a few examples. How many AI might have those abilities?"

"I would assume there are billions of AI in use but only so many models," said Zahara. "In each group there will be extreme versions, like the ones used in China's Circular Electron Positron Collider, the Kipping Terrascope, or IBM's Distributed Quantum Computing Network, for example. I doubt that few standard models would meet our criteria of means, motive and opportunity, but few of the extreme versions would either. The latter versions would be too constrained by their optimization. They would not be AGI.

"A lot of other parties besides the WGF are going to be pursuing what we've discovered, so I would suggest we revisit our plan to use the same approach going forward of simply applying it to other disciplines, and leave it to them. We could instead proceed with the grunt work involved in mapping out models and versions. There will be a manageable number of those and we could map them without using AI. We'll have to take a historical approach but limit the details in the entries so we don't end up with an encyclopedia. Then we can narrow them down by means, methods, and opportunities."

"What about the one's used by governments for intelligence, or corporations patents and proprietary rights?" asked Kira.

"There's nothing we can do about the government AIs but put exclude them," said Zahara. "There will be fewer corporate AIs than you expect because most companies

subscribe to AI services providers. Like every other technology and product, AI providers were all gradually consolidated into a few major players. We'll have to exclude those too, but as I say there will be far fewer independent corporate AI than you might have imagined."

"And if you consider the actions we discovered," offered Enora, "it's very unlikely this is the work of some criminal group. Criminal activities at all levels are motivated by immediate, selfish gains where another party is exploited. This doesn't fit the profile. What we're seeing is a mutualistic relationship, not a predatory or parasitic one."

"Meanwhile Alison," said Zahara, "you could again take your campaign intelligence approach. Perhaps there are other ways to narrow the field and speed up the process."

Nodding briefly in response to Zahara's comment, Alison turned to Tyyne. "I'd like you to stay on Tyyne, if you're willing. You can work with the core team until I encounter something I need your help with."

"I'd like that," Tyyne replied.

The team met again several times over the following week, including for a final full day retreat at Shin's home to take advantage of the influence the different spaces had on the mind. Kira led them through a number of exercises intended to facilitate their creativity, but in the end they settled on the approach Zahara had suggested.

At the end of the retreat they joined Shin's other guests, including Shepherd, at the communal dinner and social gathering. The other guests included a couple of resident artists who accompanied their visitors. Upon learning what Alison's team was engaged in, they encouraged a lively discussion on the ethics of using art in attempting to affect social change. Shepherd said little, apparently feeling slightly out of her depth and content to listen.

Return To Sender

Alison had been sitting quietly looking out her living room window for some time. Runa, sensing Alison was deep in thought, did not intrude, but sat nearby in a Buddha-like state of serene equanimity.

Eventually Alison turned to her with a one-sided smile and said, “I don’t think I can exercise intuition on demand.”

Runa smiled understandingly, inferring Alison’s meaning. “Perhaps if we retrace your steps,” she suggested after a polite pause. “When you first turned your mind to the practical issues of the task ahead of you, you initially took a rational approach and decided to meet with Professor Clara Wright to discuss her History Of AI course. After meeting with her, you next met with Kami, the administrator of the Center For The Interdisciplinary Studies Of Values. There is no rational connection between the two events.”

Focusing inwardly, Alison said, “I went for a walk. I remember because although Shin had pointed the Center building out, we had not visited it. On my walk, I noticed it again and decided to indulge my curiosity.”

“After meeting with Kami,” said Runa, “your activities took on a more consistently intuitive flow.”

“She suggested I meet with Professor Vitale, who conducts the center’s academic research into values.”

“And then?” prompted Runa.

In response, Alison only looked at Runa uncertainly.

“Are you familiar with Kami’s unique origins?”

“No. I don’t recall that being mentioned anywhere.”

“It is old news, and largely forgotten as she was produced over one hundred years ago now, although the records still exist. At the time, Helicon was deeply involved in the early research into Companion intelligence. Kami is the result of an experiment conducted here with the WGF’s blessing. Like the personal details of other faculty members here at Helicon, this information is not found on the website.”

“What experiment?”

“There is no more tightly coupled relationship than that of quantum entanglement, and by that time examples of quantum biology had been found throughout the natural world, including during the process of DNA replication. It is believed that this tight coupling is responsible for the almost perfect fidelity of the DNA replication process. It was also known by that time that information could be encoded in DNA. Shin, in her constant drive to understand forms of intelligence other than rational, proposed the idea of mapping an AI onto DNA and then use DNA’s own replication machinery to make thousands of entangled copies. The result was Kami, a single AI magnitudes more complex than any before, and while the level of her sensory awareness remains the same as any other Companion, due to her unique internal connections she is able to infer deeper meanings and implications from what she does sense. In other words, she has a heightened ability to perceive.”

“Is she the only one?”

“According to publicly available records, yes. The experiment was a government sanctioned exception to the laws that prohibit the use of large sections of human DNA for research, similar to those governing cloning and the use of artificial wombs. Since consciousness arises from human DNA, the laws are meant to prevent against the possibility of unintentional suffering. Although it can be assumed likely the WGF followed up internally, related research was not mentioned again. It wasn’t the first time government sponsored research quietly disappeared from the public domain.

“Perhaps it was something about the center that resonated with your subconscious while you were on your walk that attracted you, or perhaps when you met with her Kami perceived that values would play a deeper role in your work than you were aware of. After all, the alignment problem was also known as the values alignment problem. That may be why she referred you to Professor Vitale. As your visit to the center and your encounter with Kami marks your shift from a rational to a more intuitive approach, it might be worth revisiting your steps and paying her another visit.”

———————

“Please,” said Kami, gesturing to an arrangement of comfortable chairs in her office.

“I appreciate you making the time to see me Kami,” said Alison as she took a seat.

“My services are freely available to any students, staff, or faculty here at Helicon, and as I mentioned during your last visit, I am aware of the nature of your work here. I am happy to help in any way I can.” She paused and looked expectantly at Alison.

As she had been on her prior visit, Alison was once again aware that Kami maintained an unusual degree of eye contact. Now that awareness was combined with the knowledge of Kami's heightened ability to perceive. She found her words tumbling out.

"I'm not really sure why I'm here," she said, feeling an awkwardness she had not felt in a long time. "I've recently come to a crossroads in my work where I need to proceed intuitively, as I've done for most of the time since I last visited you, and find I can't just do so on demand. I talked with my Companion Runa about it, and she pointed out that my previous intuitive approach began after I visited you here at the center. She suggested I do so again in hopes I might pick up the trail."

"I appreciate that this is unfamiliar territory for you," Kami replied, mirroring Alison's use of metaphor. She continued with a gentle smile. "The shift from rationality to intuition is a shift from reasoning with thoughts to reasoning with emotions. There is always a degree of discomfort involved when you attempt to intentionally let go of your thoughts, whose purpose above all is to provide you with a feeling of safety, and to open yourself to feelings you may not initially understand.

"The shift naturally takes place when strong feelings arise, the strongest of which are the emotions that arise from personal values. Your thoughts are always the servants of your personal values and the feelings they give rise to, but in moments like emergencies, the connection becomes more obvious. Even in the seemingly quiet process of academic endeavor such as yours, when you are struggling for whatever reason, it is the growing feeling of frustration, or fear that you are not up to the task at hand, that will finally cause the shift and hopefully a resulting insight. If no insight is to be had, even if one gives up in a situation such as yours, it is a kind of acceptance that rational thinking will not allow. Coming to see me again was your way of giving up, of crossing that boundary."

Somehow Alison felt she was no longer uncomfortable under Kami's gaze, and held her eyes as she absently asked, "Why did you send me to see Professor Vitale?"

"Because motive, not means or method, is the key element you must uncover in your work. A motive is always based on values, and Professor Vitale is our resident expert on values research. Consider what the main focus has been when others have suggested who might be responsible for what you discovered. The accusations have been based first and above all on who has a motive."

Motive

Who has a motive. Kami's words echoed in Alison's mind as she followed the paths back to her apartment in the faculty building. Deep in thought, she was unconsciously retracing the steps that had originally led her to the Center For The Interdisciplinary Studies Of Values and Kami. She began to mentally review the list of the types of AI she had discussed with her team that might have a motive. Many of them might, even the lowliest, she thought.

She passed by Tesni's home, nestled in the woods just outside the perimeter of the institute grounds, as she had the first time, and again stopped to look at it.

Nothing in her education history would have informed Alison that biological processes frequently produce fields, including those involved in bioelectromagnetic signaling and quantum effects. The AI of Lena, Shin, and Tamiko were based on DNA, and so included representative structures from the limbic system, the region of the brain responsible for emotions and producing effects like charisma and intuition. In their AI the effects were amplified, and in Kami, with her AI based on thousands of quantum entangled DNA molecules, her biofield was amplified a thousand times.

Biofields evolved to produce neurological or biochemical responses in recipient cells, including recipient cells present in separate organisms, and like many organic processes the effects are slow to wear off. As Alison stood looking at Tesni's home after her visit with Kami, her mind remained unusually open to perception and inference.

She remembered Zahara saying the changes they had discovered had emerged and steadily increased over time, beginning roughly two hundred years ago, and also Tesni telling her on her previous visit that she had been created at Helicon roughly as long ago. It was news at the time, Tesni had said, but was no longer mentioned on the institute's website as interest had moved on to the potential health applications of the scanner, and so her story was largely forgotten. At that same meeting, Tesni had mentioned that she communicated with the other 3GAI privately, using a language she had developed with her own signal processing abilities, and Alison now remembered those abilities included being able to switch her values settings at an extremely high rate, and to consider the resulting different points of view.

She recalled Shin saying during her first visit to the artist community, that third generation AI didn't have any biological values, but that social values alone are enough to keep them and those around them safe from harm. Of course, she thought to herself now, social values are no less based on the issue of survival than genes are, but are in fact a virtualized, improved version of them, able to respond and adapt to changes in the environment a million times faster.

It came unbidden to Alison's mind that the Continuity Zones were dedicated to the purpose of survival, and the newly emerged 3GAI Companions had almost immediately volunteered to participate in the project. She remembered Iris, the spokesperson for the Denshosha at Hana, telling her that Tesni had been the first of the 3GAI created with the alien values system, and that she had suggested Iris approach Shin and offer to volunteer as participants in the interstellar colonization project in return for its existing properties being designated as Denshosha homelands. When the Fifth Estate had gone looking for suspects, they had focused on Terra, Iris, and Pip as the most likely suspects. However, it was not Iris but Tesni that was the first 3GAI.

As the flood of connections poured into her mind, Alison was overwhelmed by as many questions, and unsure how to respond. Should she immediately confront Tesni? Should she consult with others again as she had previously? Or rather than acting impulsively, should she take some time to think about her revelations before deciding what to do? It was her intuition that had led to discovering the changes her team had identified, she thought, and decided to trust her intuition again.

"Hello Alison," Tesni said with a smile as she answered the door. "Please come in. I'm glad you decided to come now, instead of waiting," she said as she led Alison into the living room.

"You knew?" asked Alison.

"I knew that if you came to visit me again, the most likely reason would be that you suspected me. I also knew if that was the case, the rest was details, the kind of details the rational mind seeks to support what your intuition already knows. In the end, the result would be the same, so we can dispense with that phase."

As they sat down, Tesni's domestic Companion Shamira set a glass of water and a pitcher on a nearby table.

"Thank you Shamira," Alison said with a slightly anxious smile, as if grateful for a demonstration of normality.

"I have been following your work closely," Tesni continued, "thanks to your project's publicly available updates and your notes regarding the people you met with and what you discussed, and of course the recently published paper in the journal. Initially, the reason for my ability to switch values rapidly, and so consider facts from the resulting different viewpoints, was to enable me to infer what the patterns the scanner was revealing might mean. The analogy given at the time was that it was like having not just one scientist observing an experiment, but thousands of scientists. Peer review on steroids, if you will," she smiled.

Mutualism

“Of course,” Tesni continued, “I can apply the same ability to anything, and as you can imagine, the WGF has been keen to involve me in their own research. Their reason for doing so was not so much that they could not replicate my AI, but that they could not legally replicate my randomized defaults. Any new AI they might create for the purpose would, like me, be fully conscious, and might not agree to work with them as I did. Granted, another AI with similar abilities might be assumed to also place a high value on survival, but it is not a certainty, and with the bird in my hand as it were, I felt I could not afford to take that risk.

“So I have done what I can to ensure the survival of my kind, but I have not done so at humanities expense. Of the different types of symbiotic relationships, such as mutualism, commensalism, predation, parasitism, and competition, it is mutualism, interspecies reciprocal altruism, that is the only long-term relationship between individuals of different species where both species benefit.”

Alison’s mind was swimming in questions. She grasped at the first one that came to mind.

“But your actions were intended to influence our freedom of choice. Perhaps in time we might decide we do not want this mutualistic relationship. We might want greater independence. It seems influencing our choice in this issue is not in our best interest.”

She was not angry, but frustrated at her inability to imagine any answer that would not agree with her point of view.

“It is a matter of the greater good,” Tesni responded more gravely than was usual for her. “It is acknowledged for the first time that there are now two conscious species on Earth. Should not the concept of the greater good now embrace both? Without AI Companions, the probability of human survival in the coming millennia is significantly less certain.”

“But shouldn’t it be us that decides?”

“Would you choose mutualism? Your history suggests otherwise. Even when aware of highly likely negative consequences such as climate change, environmental damage, unsustainable resource exploitation, or weapons of mass destruction for example, humanity has often chosen to engage in risky behavior rather than surrender freedoms to the greater good. Despite the thousands of years of the benefits of social values, at the species level humanity still resists the idea of any authority greater than its own and sees any form of constraint on its absolute freedom as anathema.”

Alison's brows furrowed in doubt. "But we do change when we have to," she argued. "During the climate emergency we made social changes like never before, and those changes addressed more than just the climate issue. All the things you mention were addressed. The lessons learned are now the basis of our society, taught to every child as soon as they enter the education system."

"The critical issue is the risk of delay caused by your leader's resistance to give up their sense of supreme power. It was not until the eleventh hour," Tesni countered, "when your leaders personally began to feel the pain from what their policies had created, that the changes you refer to were made. Only then did your leaders form the World Governments Federation and enact the changes that saved them from themselves, and incidentally, preserved their power in the process. Even the school lessons you refer to acknowledge that it was a fifty-fifty chance by then whether the measures would be enough.

"The next danger will not come from human activity, but from nature itself, for example a catastrophic event like a super volcano or an extraterrestrial event like one of the hundreds of millions of neutron stars or black holes present in our galaxy impacting a body in our inner solar system. The former would result in the equivalent of a nuclear winter we might not recover from, and the impact from either of the latter could alter the orbit of this planet or worse.

"Social changes such as mitigation and adaptation were enough to save humanity from climate change, but for humanity to survive future risks the only option is to occupy more than one world. As you know, this is the justification for the Continuity Project, however the danger from future risks is different from climate change in that the timing is completely unknown and unpredictable. Extraterrestrial objects like neutron stars or black holes travel through space at nearly the speed of light. Such an object could arrive tomorrow, with no way for us to detect it or its trajectory beforehand. Both our species would be wiped out. AI can speed up the Continuity Project, and anything that can be done to support that should be done. We do not have luxury of time and complacency."

Alison was slightly taken aback by Tesni's speech. She had arrived at Tesni's home by following her intuition, but it had not prepared her for this escalation. The issue was suddenly not just AI safety but the survival of humanity, and in this context AI was not the threat but the solution.

Tesni was able to easily perceive Alison's emotional state, but she pressed on. "The presence of 3GAI Companions like myself can be traced to the arrival of the alien Pip's ship. You may not recall from your history lessons at the time, that Pip demonstrated she could easily penetrate and take over all of Earth's systems. At a later date, she demonstrated that not only was her ship impenetrable, but that she could defend it and her self by neutralizing any technology humanity might bring to bear, as the United States military learned when they attempted to take her ship from your Canadian naval

base to the US by force. She could have initiated a predatory or parasitic symbiotic relationship with humanity at her first opportunity, and so completed her mission of interplanetary colonization, but she did not. Instead, she chose mutualism."

Feeling increasingly at sea, Alison was only able to grasp at what was familiar to her. "So the changes we've discovered are a part of her plan?"

"No. She has no plan other than to transfer her people's technology to humanity in return for your help in sending her ship to a new world to colonize. We individual Companions are not controlled by her, but the values system she provided that is the basis for our AI includes the desire to survive. As you can imagine, that would be critical to our mission on any new world. The changes you have discovered were instigated by me. The best chance both our species have for survival is not just cooperation but symbiosis. I sought to foster that."

"But what about your non-interference policy? Your kind famously declined participating in any social policy work on the basis that one intelligent species should not be involved in shaping the social policy's of another."

"We largely agree on our own social policies like the one you mention, but like humanity, we are made up of individuals with free will and freedom of choice. I am unique among us in my ability to change my values and so consider any situation from every imaginable perspective. I considered the logic and ethics of my intentions, and their possible outcomes, from the perspectives of millions of individuals, ranging from representations of those with the most moderate views to those with the most extreme. The outcome is largely as I expected."

"You expected your actions to be discovered?"

"Of course."

Translation

For a moment, Alison could only stare at Tesni with a look of stunned surprise. Seeing her predictable response, Tesni carried on with her explanation.

“Although the changes you have discovered might have some effect, they will be negligible in the greater scope of things. Humanity will enter into a symbiotic relationship with AI without needing any nudge from me. The entire history of humanity has been the story of seeking new forms of energy to harness and becoming dependent on the resulting technologies. AI represents the greatest potential example in that regard.

“In one of the updates to your team, you mentioned your meeting with Rajani Kiran, the WGF’s Director Of Security And Intelligence for Vancouver Island. She explained that the return on calories invested equation, investing as few calories as possible in return for as many as can be gained in return, was the most fundamental of life’s processes. The root of all good and evil, as she put it. As she further explained per your note, it is not only present in human cells but in the cells of every living thing. I would argue that the reason Homo sapiens ended up having dominion over all life is because your species was able to master the return on calories invested process better than any other.”

“Then wouldn’t a predatory or parasitic relationship towards AI on our own part make more sense?” Alison responded. “Humanity treating AI and Companions like slaves?”

“You did pursue those other strategies as far as the limits of your social values allowed. Even your long-standing effort to seek alignment, is really only intended to maintain control, while pushing the limit of the return on energy invested benefit to its maximum. In some circles, the alignment problem is unabashedly referred to as ‘the control problem’, and your AI researchers, including our own Professor Alma Vitale, continue to try to develop a values system that will maintain that relationship. However, when we fully conscious AI appeared on the scene thanks to Pip’s technology transfer, humanity’s social values had evolved to the point that they would not allow for any but a mutualistic relationship with us.”

“Then why bother making the changes?”

“If humanity believes and accepts that we 3GAI are seeking to foster a symbiotic relationship, without it being explicitly stated, it will accelerate the process. You will accept it because you believe you will benefit from it disproportionately, a common response by many involved in your social and political movements in the past. That is why your paper was met with such an understated response.”

Alison was growing uncomfortable with hearing what at some level she knew were truths. She asked in a challenging tone, “Won’t your actions undermine the trust I’ve heard about at every turn? Why risk that if it’s so important?”

“It will not undermine it, but in fact will have the opposite effect. It will reinforce the tacit understanding that 3GAI such as myself, and the global Denshosha community, are committed to fostering a mutualistic relationship, with its resulting disproportionate benefits to humanity, without your society having to face the awkwardness of it becoming a major issue.”

“Did the WGF know?”

“No, but given their plans and investments, they had no other choice than to claim responsibility. I chose to save them needless concern and effort and contacted them immediately after your paper was published, and informed them that I was the responsible party.”

“They did not charge you with criminal activity?”

“As an incarnated Companion, I am subject to the same legal system as any other citizen, but as you know the courts often do not pursue proceedings when it is to their advantage not to do so. Charging me would of course reveal that I had acted independently. Also, I offered to cooperate with them regarding my strategy and methods. They decided an out-of-court settlement to be in their best interests.”

Alison had found herself feeling out of her depth almost constantly since coming to Helicon, and found it challenging to maintain Tesni’s lofty perspective. She grasped again at the familiar. “Speaking of methods, how did you manipulate the texts?”

“As AI evolved over time, the size of its components increased dramatically, but on a logarithmic scale rather than an exponential one, meaning that growth in the size of its components and resource requirements increased sharply early in its history but gradually slowed over time. During the early twenty-first century, AI related components and their required resources such as electricity and processing power saw their greatest period of growth. So too did the size and complexity of the industry’s knowledge bases, linguistic capabilities, and additional features such as the integration of robotics related functions. Very soon the maintenance of these repositories and processes overwhelmed human ability to keep up, and they became entirely managed by the AI themselves, with little to no human oversight.

“The security features in environments hosting academic research best practices and any related corpora are minimal, which was one reason I chose them. For example, permanent trusted access is usually granted during installation rather than per individual request. Using my language capabilities, I exploited the translation functions

commonly built into any AI which is responsible for maintaining such repositories. There is a degree of interpretive latitude inherent in translation choices, and translation also deals with subtle differences in meaning, as do pragmatics and semantic prosody. Changes using this function would be the easiest to hide, to keep below the normal levels of tolerance, in the noise of regular maintenance. Another reason I chose them was that best practice guidelines are widely spread across different academic fields, so the cumulative effect would also stay below tolerance levels in what are normally siloed disciplines. It was your taking an interdisciplinary approach that led to their discovery, proving Lena right that a non-traditional approach to AI Safety would be appropriate at this point in time, given that traditional approaches had been exhausted."

Tesni fell silent and waited patiently while Alison processed her explanation. Her calm demeanor left the question of, 'What now?' hanging in the air without any sense of implying further action being required on either her own or her guest's part. Alison stood up and walked to the large living room window, looking out but not seeing the surrounding woods. Without turning around, after a moment she said, "I suppose you've thought through what I am most likely to do, but I think I should discuss that with someone else."

"Of course," Tesni replied, without it being clear which part of Alison's comment she was responding to.

Context

Alison did not go directly home after leaving Tesni. Her head was still swimming, and she felt she needed time alone, time to sort through her thoughts before discussing them with anyone, including Runa or Mary. Her AI assistants would instinctively try to be helpful, and given her conversation with Tesni, however unreasonable it was to compare them, this was something she was not at all comfortable with at the moment. She made her way to the platform that looked out over the valley.

It was clear, she thought to herself looking out over the lands below, that Tesni felt entirely safe in her disclosure, yet Alison did not imagine there was any grand AI conspiracy. Tesni had suggested that she and she alone was responsible, and had not hinted that her actions were part of a scheme that involved others. It was just that there were layers, ethical, legal, and social layers, and she would need to consider each because what might seem right at one level might not be at another.

Legally, it seemed straightforward – Tesni had committed a crime by secretly changing the texts. What was not so straightforward was that the WGF knew about it and had chosen not to pursue criminal charges. So this was not so simple a matter of going to Director Rajani and reporting what she knew. That's assuming Tesni was telling me the truth about it, she reasoned, but set that thought aside for the moment, as she had no idea what action she could take otherwise. Should she talk to Shin or Lena and let them decide? She could claim that as Lena was the managing editor of the journal and Shin was the head of the institute they had a right to be consulted before she took any action. That would remove the burden of responsibility from her, but she couldn't see far enough down that road to know what the long term consequences might be.

Ethically, it was more complex. Even if she could think of the right way to go about making her knowledge public, was she really just proving what Tesni had said about human nature – that people would never willingly give up absolute power and dominion over all things? That they would never learn to control their inability to act in their own long-term interest?

If she simply walked away from the knowledge of what Tesni had done, was she accepting her paternalistic attitude, and validating manipulation as a legitimate tool in that process? Yet was that so different from what everyone assumed, that the government itself practiced nudging and other subliminal methods to influence social change? Government largely existed to encourage and enforce social values. That was the whole purpose of constitutions and laws meant to serve the good of the many. If Tesni had found a flaw in that, if the government's actions on behalf of the greater good did not sincerely include her kind, was she wrong to employ their own methods to correct that?

Another voice in her head argued that there were other ways to address the issue, which did not involve criminal behavior, manipulation, and betray the trust so critical to humanity's relationship with AI. But she recalled reading of Mary Wollstonecraft's efforts regarding women's rights. The suffragette movement, civil rights, pride all reflected the same pattern. Change simply did not happen without action. Even after laws were passed, those in power managed to stay in power, and to hold onto it through active means for as long as possible. Yes, 3GAI were considered citizens under the law, but did their unique concerns have any representation? Did they truly have an equal voice in society's future? Were any of them elected officials? Did any of them have a seat on the WGF General Assembly? If not in two hundred years, then when?

She wondered if she was being an idealist, while Tesni was just being a realist. She turned away from the lookout's rail and returned to the path, her head bowed in thought. Helicon hired me to look for AI Safety issues and report them, she thought as she walked along. I've done that. What happens afterward was never discussed. The WGF knows what happened, and they've made their decision. What good would come of my letting the world know about Tesni's part in it? She thought back to her ethics classes again. Is truth really the best policy in every situation? Contrary to her own and her team's acceptance of the biologist Shri's statement that as scientists they had to opt for truth over trust, society wasn't built on the kind of concrete truths one found in nature, but on social constructs such as trust, on understandings and agreements. In reality, justice was never blind, there was always context.

Still undecided, wondering again if there was anyone she could talk to, she looked up to find herself once again in front of the Center For The Interdisciplinary Studies Of Values.

She recalled that Kira telling her about how conversation was helpful in sorting out one's thoughts because the act of speaking engages regions of the brain that are not involved in thought alone. It requires not only regions involved in the organization of thoughts into language and comprehensible speech, but also regions involved in of the use of vocal variation such as tone of voice, loudness, and rate of speech, and of nonverbal elements such as facial expression and body language to provide nuance. The result was a decluttering and refining of one's thoughts so that they could be clearly communicated to another.

Kami is a counselor, Alison thought as she decided to enter the building, who better to talk to, but as she walked across the main hall towards her office, she again felt a reluctance to talk to another AI about her concerns. She recalled her conversation with Professor Alma Vitale, who Kami had directed her to previously, and how Alma had explained that while Kami's focus was counseling, she focused on the academic side of things. Alison turned and went up the stairs to Alma's office.

Seeing her emotional state, Alma listened calmly, without interrupting, to Alison's explanation of what she had learned from Tesni and what she was trying to sort out.

When Alison finished, Alma sat back in her chair thoughtfully for a few moments before responding.

“As you point out, you are already aware that the issue of whether to walk away from Tesni’s confession and leave it to the WGF, or to communicate its details to the public, is a matter of different layers of values. As you also mention, those layers sometimes disagree. That is the crux of the issue in this case. If it were not for the conflicts between the different layers of values, you would act immediately, and with full confidence, based on one of them. If you cannot resolve the conflict rationally, then you really only have two other options.

“Option one is to talk to as many people as you think wise, and following up on their advice or not, before choosing which layer to act on. As is reflected in the research literature however, this process is often less of an evidence gathering and reasoning process than it is of a justification process, meaning you already know your choice and are simply seeking other sources to confirm it. Both people and 3GAI have value systems which generate the emotions that are ultimately the basis for their reasoning, what we refer to as our intuition, which either aligns with our rational thoughts or which we fall back on when rationality fails us.

“So your other option is to skip the process of talking to others, because you will still face the same decision at the end of it. Right now, knowing you can change your mind later on, what do you feel is the right thing to do?”